DOCUMENT RESUME

ED 255 545 TM 850 145

AUTHOR Braun, Henry I.; Jones, Douglas H.
TITLE Use of Empirical Bayes Methods in the Study of the Validity of Academic Predictors of Graduate School Performance.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-84-34; GREB-79-13P
PUB DATE Feb 85
NOTE 82p.
PUB TYPE Reports - Research/Technical (143)

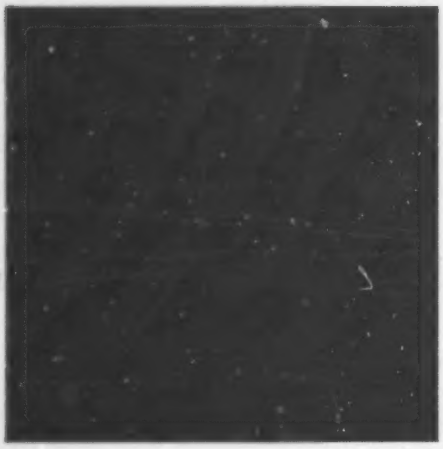
EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *College Entrance Examinations; Departments; Grade Point Average; *Grade Prediction; *Graduate Students; Higher Education; Least Squares Statistics; Mathematical Models; Multiple Regression Analysis; *Predictive Validity; Predictor Variables
IDENTIFIERS *Graduate Record Examinations; Graduate Record Exam Validity Study Service

ABSTRACT

Classical statistical methods and the small enrollments in graduate departments have constrained the Graduate Record Examinations (GRE) Validity Study Service to providing only validities for single predictors. Estimates of the validity of two or more predictors, used jointly, are considered too unreliable because the corresponding prediction equations often possess implausible characteristics. This study investigates two statistical methods--empirical Bayes and cluster analysis--to determine their applicability to these validity problems. Data on 6,946 students from 190 participating departments were used. It is concluded that, by using the new class of empirical Bayes methods, it is possible to obtain, at the department level, useful and reliable estimates of the joint validity of several predictors of academic performance. Further methodological refinement will allow the question of differential predictive validity to be addressed at the departmental level. The technical appendix describes the estimation problem following from the empirical Bayes model. (BS)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED255545



GRE

GRADUATE RECORD EXAMINATIONS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

N. Weideman Miller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

# USE OF EMPIRICAL BAYES METHODS IN THE STUDY OF THE VALIDITY OF ACADEMIC PREDICTORS OF GRADUATE SCHOOL PERFORMANCE

Henry I. Braun
Douglas H. Jones

GRE Board Professional Report GREB No. 79-13p
ETS Research Report 84-34

February 1985

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

TM 850 183

# Use of Empirical Bayes Methods in the Study of the Validity of Academic Predictors of Graduate School Performance

Henry I. Braun

and

Douglas H. Jones

GRE Board Professional Report GREB No. 79-13

February 1985

This report is not to be quoted without consent of the authors.

Copyright, © 1985 by Educational Testing Service. All rights reserved.

Acknowledgements: The authors would like to thank Donald B. Rubin and Paul W. Holland for helpful conversations; Bruce Kaplan and Dorothy Thayer for programming support; Nancy Burton, Dorothy Thayer, and Ken Wilson for comments on our earlier draft; and Linda DeLauro and Vera House for assistance in the preparation of this report.

## Executive Summary

Graduate education in the United States is characterized by an enormous diversity of disciplines and the predominance of relatively small enrollments in individual departments. In this setting, a validity study based on a single department's data and employing classical statistical methods can be of only limited utility and applicability. At present, to participate in the Graduate Record Examinations Validity Study Service, a department must have at least 25 students in its entering class. Only validities for single predictors are provided; estimates of the validity of two or more predictors, used jointly, are considered too unreliable because the corresponding prediction equations often possess implausible characteristics, such as negative coefficients. These constraints were introduced by the Validity Study Service to reduce the chance that the results in the report to a department would be overly influenced by statistical artifacts in the data and hence serve more to mislead than to inform. Two unfortunate consequences, however, are that fewer departments than before can benefit from the service and those that do cannct obtain information on a number of issues of interest. For example, questions of the incremental contribution to validity of one predictor when two or more predictors are already included in a prediction formula cannot be answered. Moreover, even the single predictor validities display considerable year-to-year fluctuation, which may bewilder departments that regularly receive Validity Study Service reports.

Until recently, little could be done to ameliorate this unsatisfactory state of affairs. Although a number of strategies for pooling data across departments have been proposed, they all suffer from at least one drawback: Either the pooling is so indiscriminate that the results appear to have questionable relevance to an individual department or it is so delicate that it would appear to be difficult to implement in an operational setting.

The goal of the present study was to investigate two statistical methods, empirical Bayes and cluster analysis, to determine whether their application to the problems faced by the Validity Study Service could result in useful improvements. Because of the successful application of empirical Bayes methods to validity problems in another context, particular emphasis was placed on this approach. In fact, considerable effort was expended in developing and studying a new and more general class of empirical Bayes models that can accommodate the complex structure of the Validity Study Service data base.

To borrow a term from sociology, empirical Bayes facilitates a very general form of "contextual analysis" of the validity problem. Essentially, the relation between the criterion and a constellation of predictors within a given department is examined in the setting of a large collection of departments. Of particular interest is any evidence that the nature of this relation varies in association with some measured characteristic(s) of the departments. An example might be the finding that the inclination of the regression plane increases as the department size increases. To the extent that such pandepartmental findings are valid, the precision of the estimation

carried out in any one department can be improved by drawing upon the information provided by the other departments.

Two technical points must be satisfactorily addressed before such a procedure can be implemented. The first is to determine how much the estimates based on a single department's data should be modified by data from other departments. The empirical Bayes methodology provides a good solution that depends on both the precision of the within-department estimate and the apparent strength of the pandepartmental relation. Details are given in the text. The second point concerns which departments are to be considered together. Various possibilities exist, including the formation of clusters of departments on the basis of either substantive or statistical criteria. The empirical Bayes methodology would then be applied separately to each cluster. An extreme approach is not to form clusters at all but to treat all graduate departments participating in the Validity Study Service as a single "family," trusting that certain measured departmental attributes are sufficient to characterize the departments in the validity setting.

Our analysis demonstrates that empirical Bayes does provide a useful way of combining information across departments. Interestingly, it appears to work best, in this case, when departments are characterized by various measures of student quality but are not divided into clusters determined by discipline, location on a verbal-quantitative axis, or various other statistical criteria. The practical result is that, even for departments with as few as 10 students, separate

prediction equations have been obtained from which stable estimates of the joint validity of two or more predictors can be derived.

An important component of the study was the comparison of various suggested approaches along a number of dimensions, especially those related to the quality of their predictions. The quantitative nature of these comparisons may seem somewhat at odds with the way in which most graduate departments probably utilize the results of a validity study. Rather than using the prediction equations provided to make exact predictions of first-year averages to be earned by prospective candidates, they look for guidance on the relative weights to be assigned to various predictors in making a qualitative assessment of the candidates. However, in our view, there can be little justification in proposing that empirical Bayes methods supplant classical least squares unless it can be shown that, among other things, the change would result in prediction equations that yield better predictions and are more stable through time.

Among the approaches compared were ordinary least squares, least squares in conjunction with pooling of data, and a variety of empirical Bayes methods in conjunction with different levels of clustering of departments. To simulate the admissions setting, most of the comparisons were carried out through cross-validation: Each department was randomly divided in half; models were estimated with one half-sample, the calibration sample, and tested on the other half, the validation sample. The comparisons demonstrated that a fairly simple empirical Bayes model not only yielded better predictions of first-year averages but also facilitated the accurate

assessment, a priori, of the quality of these predictions. Moreover, the prediction equations were quite stable and rarely displayed implausible features such as negative weights.

How do our results compare with those reported by Livingston and Turner (1982)? Although most of their report centers on zero-order correlations, they do report joint validities for verbal and quantitative scores and undergraduate grade point average, estimated either by pooling data across disciplines or by forcing the predictors to be equally weighted. These validities fall in the range of 0.25 to 0.45 and are somewhat lower than the validities that can be realized using prediction equations obtained through empirical Bayes. These range from 0.30 to 0.55.

In another phase of the research program, a new statistical method for hierarchical clustering was developed and applied to applicants' characteristics. These characteristics include performance on the GRE General Test and the Subject Test taken. Both the resulting clusters and the original candidate information were employed in a number of ways in estimating validity, but the results did not offer an improvement over the empirical Bayes model. Nonetheless, the new clustering of disciplines may be of some interest in its own right.

The empirical Bayes method has also been adapted to the problem of providing separate estimates of validity for various subgroups of a given population of students. Two trials were carried out: one in which students were categorized by race and another in which they were categorized by both age and sex. In both cases, the differences

in validity were neither of practical importance nor statistically significant. However, the relatively small amount of data on minorities suggests that real differences, if they existed, would be difficult to uncover.

The principal methodological conclusions of this study are that, through the use of a new class of empirical Bayes methods, it is possible to obtain, at the departmental level, useful and reliable estimates of the joint validity of several predictors of academic performance and that these methods may be further refined to address the question of differential predictive validity, again at the departmental level. These results have important practical implications for the GRE Validity Study Service.

## 1. Introduction

The research described in this report was undertaken both to develop solutions to some basic problems that have plagued validity studies of Graduate Record Examinations (GRE) scores and to broaden the range of questions the GRE Validity Study Service might usefully address.

To examine the validity of a single measure as a predictor of first-year average (FYA) in a graduate department, the standard approach has been to carry out a simple regression of the criterion, FYA, on the measure, based on data culled from a recent cohort of students successfully completing their first year. Similarly, a multiple regression must be performed to assess the joint validity of two or more predictors. In either case, the appropriate validity (or correlation) coefficients may be easily derived from the fitted regressions.

Unfortunately, these validity estimates are affected by the very nature of the process by which students are admitted to graduate school. In particular, if the predictors whose validity is to be assessed are employed in the selection process, the attending students will display a distribution of test score values that differs from that of the applicants. For example, the proportion of attending students with low test scores will usually be smaller than that of applicants. The corresponding validity estimates will tend to be lower than they would otherwise be in an unselected population.

To this difficulty, encountered in virtually all validity studies, must be added the small size of most graduate departments.

Small sample sizes result in unstable estimates of validity--that is, estimates that may fluctuate wildly from year to year. In addition, the selection process will also depend on a number of unmeasured or unrecorded factors (letters of recommendation, extracurricular activities, and the like) that are not perfectly correlated with those available for study. Fluctuations in the quality of applicants on these unavailable factors across years will also contribute to the apparent instability of the fitted regressions.

Consequently, estimates of validity derived under these circumstances may be of limited utility. The present Validity Study Service has taken a number of steps to mitigate the effects of these difficulties. At one time, departments with as few as 10 students could participate in the service. Now they must have at least 25 students. Secondly, only the zero-order correlations are reported. These are the validities of each factor taken alone. The drawback to the first step is that many fewer departments may avail themselves of the service. The drawback to the second is that the joint validity of a set of predictors cannot usually be accurately inferred from their individual validity coefficients. Moreover, the zero-order correlations may be quite misleading if two or more of the factors were important elements in the selection process.

Nonetheless, the course adopted by the Validity Study Service is a wise one inasmuch as Boldt (1975) has reported that attempts to fit multiple regressions to GRE data result in numerous negative coefficients for the predictors, implying that the better the

performance on the predictor, the poorer the performance in graduate school! Such results are undoubtedly artifacts of the selection process and small sample sizes described above. A number of researchers have wrestled with these problems. Generally, their strategy has been to devise sensible ways of combining information across departments.

One of the first applications of Bayesian techniques to these problems was reported by Novick, Jackson, Thayer, and Cole (1972). Building on a series of papers by Lindley (1969, 1970), they carried out a cross-validation study on a set of American College Testing Program data. The results suggested that the Bayesian approach outperformed ordinary least squares on a within-department basis. An excellent account of related statistical work up to 1975 is provided by Boldt (1975).

Boldt compared a central prediction approach, which applies a least squares technique to College Board Validity Study Service data appropriately pooled from different sources, to the Bayes approach of Novick et al. He reported that least squares performed about as well as Bayes in a cross-validation analysis of prediction performance. Wilson (1979) used a central prediction system of common weights for departments and suggested methods for testing whether an individual department had a different prediction system. Data were not available for cross-validation of this method. Using GRE validity data, Wilson found by analysis of variance that only a small fraction of departments appeared to have different prediction systems.

Using Law School Admission Test validity data, Rubin (1980) has reported that in a cross-validation of multiple $R^2$, empirical Bayes techniques outperformed the ordinary least squares (within-school) techniques. One of the significant results was that, although each individual law school data set was large, the least squares prediction systems were shown to be unstable from year to year when compared to the empirical Bayes prediction systems. The policy of some departments in allowing high scores on one test to compensate for low scores on another, studied by Dawes (1975), has been shown to severely hamper estimation of the underlying prediction system when using a single department's data and is one cause of the instability mentioned above.

Braun and Jones (1981) have studied prediction bias in the context of the Graduate Management Admission Test for Blacks in a data set in which only 6 percent of the students were Black. Using empirical Bayes techniques based on Dempster, Rubin, and Tsutakawa (1981), they obtained prediction systems together with appropriate confidence statements for Blacks even in schools having less than five Black students. Similar techniques should be applicable to the data of the GRE Validity Study Service.

A requirement of the methods described above has been the identification of homogeneous groups of departments. Because of the diversity of graduate department types, attempts at clustering have resulted either in a large number of small groups or a small number of rather heterogeneous groups. In either case, the decision is often made in a subjective fashion. Our present approach has been to extend the

empirical Bayes methodology to accommodate the structure of the GRE Validity Study Service data base with a view to possibly dispensing with the need for such prior cluster formation. We have also considered empirical Bayes prediction systems that do utilize departmental clusters in an important way. Nonetheless, we believe that the analysis described below conclusively demonstrates that a general empirical Bayes framework is sufficient to produce a superior prediction system without recourse to any subjective grouping of departments.

A brief description of the data base we have worked with is provided in the next section. In Section 3, our empirical Bayes models are explained and compared with current procedures. The application of empirical Bayes methods to the problem of determining the predictive validity of GRE scores for various subgroups of students is addressed in Section 4. Section 5 considers alternative approaches to the clustering of departments, and the final section contains some discussion of the validity of the GRE score battery, based on inferences from the preferred model, as well as suggestions for future work.

2. Data

The information collected from participating departments by the GRE Validity Study Service from its inception through the spring of 1981 provided the essential data for this study. The information arrived in two ways. A "Prerecorded Data Collection Form" is sent by the service to each graduate department, containing the names of students who had test scores sent to the department, which then supplies for each student the year of enrollment, undergraduate grade point average (UGPA), and graduate first-year average (FYA). Using an "Add-On Data Collection Form," the department may supply information on additional students that enrolled but were not on the first list. The Validity Study Service then matches the two lists against the original score sender files, producing a single file containing score information as well as sex, ethnic status, native language, and handicap status for each student.

The original file of 8,224 students was reduced to 6,946 students from some 150 departments because of the requirement of full information on the predictor scores and criterion. In subsequent sections, when we execute cross-validation procedures or carry out analyses of differential predictive validity studies for various subgroups, small sample sizes or missing information on individual characteristics will result in further reductions in the data base employed. These reductions will be described in the appropriate sections.

In our work on clustering, we employed the "Graduate Institution Summary Statistics Reports" issued by the GRE program to all

institutions who received GRE scores for a given testing year. These reports summarize for each departmental type the distributions of GRE verbal and quantitative scores, the GRE Subject Tests taken, and the Subject Test scores of all applicants to departments of that type. The summary data are based on all scores reported between October 1978 and October 1979.

The data were scaled prior to carrying out any computations. The GRE scores were divided by 200 to make them roughly comparable to undergraduate grade point average, and the first-year averages were scaled separately for each department by subtracting out the mean and dividing by the standard deviation. Because the standard deviation can be quite variable, especially in small samples, this scaling introduces an extra measure of noise. On the other hand, the assumption of exchangeability of the regression coefficients is more plausible when the criterion scores have been standardized in this manner, and the empirical Bayes estimation procedure mitigates most of the effects of the added noise.

## 3. Empirical Bayes Models for Prediction

### 3.1 Introduction

The nature of graduate education in the United States, most notably the diversity of departments and the relatively small enrollments, has forced researchers to adopt a number of compromises in their analysis of the validity of the GRE score battery and college grades. For example, departments with fewer than 10 (or, more recently, 25) students have been excluded from studies carried out by the GRE Validity Study Service. Moreover, only zero-order correlations have been presented since it appears that the estimates of coefficients in a multiple regression are too unstable.

Another approach has been to pool data across a group of departments and to estimate a single prediction equation for the ensemble. These groups could be formed on a quasi-subjective basis (for example, all psychology departments) or on a statistical basis by testing for the equality of regressions (Wilson, 1979). Whatever the method, a number of problems have persisted, including a large proportion of negative coefficients and serious instability of estimated coefficients.

Our own efforts have been centered on applying empirical Bayes methods to the question. In a sense, empirical Bayes provides an attractive compromise between the extremes of locally determined regressions for each participating department and a common regression for a group of departments. In fact, empirical Bayes does require that a group of departments of like nature be identified. Then, under certain assumptions (specified below), separate estimates for

each department are developed. These estimates, however, are combinations of the least squares estimates based on that department's data alone and a pooled estimate based on the data from all schools in the group. The exact nature of the combination is determined by the estimated precision of the least squares estimate and the pooled estimate. Stability of the final individual estimates is enhanced by the contributions of the pooled estimates, while departmental idiosyncrasies are given some weight through the least squares estimates.

3.2 Developing Empirical Bayes Models

While the analysis to be described in succeeding sections is fairly complex, the essential notions of the empirical Bayes methodology can be illustrated by a simple example.

Suppose we wish to study by linear regression methods the relation between a criterion Y and a single predictor X in a collection of institutions, m in number. Thus, we postulate a simple model (with no intercept) of the form

$$(1) \qquad Y_{ij} = B_i X_{ij} + e_{ij} \qquad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, N_i \end{matrix}$$

where

$Y_{ij}$ = criterion score of student j in institution i

$X_{ij}$ = predictor score of student j in institution i

$e_{ij}$ = associated residual error

$B_i$ = slope of regression in institution i.

In classical statistics, the parameters $B_i$ are estimated on the basis of the data in institution i only. The particular statistics employed

in the estimation depend on the assumptions made concerning the distribution of the residual errors $e_{ij}$.

In the empirical Bayes approach, another assumption is made: namely, that the true slopes $B_i$ may be thought of as being randomly generated from some prior distribution. Although the form is specified beforehand, the parameters (called hyperparameters) of this distribution are left free to be estimated from the data. These hyperparameters, in turn, are used to provide estimates of $B_i$ that differ from those of a classical analysis. For example, we might assume

$$(2) \qquad B_i \overset{iid}{\sim} N(\mu^*, \Sigma^*)$$

where $\mu^*$ and $\Sigma^*$ are the free scalar hyperparameters that must be estimated from the data.

The key implication of the empirical Bayes assumption is that the data in one set of institutions contain information about the value of the slope in each individual institution. Thus, a mechanism is provided whereby the information in the entire data set can be employed in the estimation of the individual slope parameter. The practical consequence, in the present context, is that the empirical Bayes estimate is derived by pulling the individual estimate towards the estimate derived by pooling the data over institutions. The greater the estimated imprecision in the individual estimate, the more it is modified.

In mathematical terminology, the proper application of empirical Bayes methods depends on the assumption that the units of the analysis (in this case, the departmental regression coefficients) constitute

an exchangeable group of units. Assumption (2) is one very special way of formulating this assumption. Essentially it means that there is no reason to believe, a priori, that the value of the regression coefficients for one department are larger or smaller than those for another department.

Most researchers, we believe, would be quite comfortable with the assumption of exchangeability for a group of like departments, such as a group of anthropology departments, known to be equally competitive and with similar grading standards. However, if we wished to treat simultaneously a more heterogeneous group of departments, such as all physical science departments, exchangeability or, in particular, assumption (2) would no longer appear very plausible. Nevertheless, by generalizing (2), it is possible to remain in the empirical Bayes framework.

For example, imagine that a set of the regression coefficients is not exchangeable because their values vary systematically with some other departmental factor, Z. That is, suppose that in place of (2), we have

$$B_i = Z_i G + D_i \tag{3}$$

where G is an unknown coefficient, and D is a random error for which it is assumed that

$$D_i \overset{iid}{\sim} N(\mu, \Sigma^*). \tag{4}$$

Just as in ordinary multiple regression, the coefficient G determines how the departmental covariate must be weighted to yield the regression coefficient $B_i$. Thus the only novelty is the nature of the criterion.

A convenient choice for $Z_i$ is $X_{i.}$, the mean score on the predictor X among students in the department. Note that we are assuming exchangeability for the errors $D_i$ about the regression. In effect, we have retained exchangeability in the model by explicitly modeling the systematic (nonexchangeable) part of the distribution of the regression coefficients. In principle, this model can be tested by embedding it in a still more complex system.

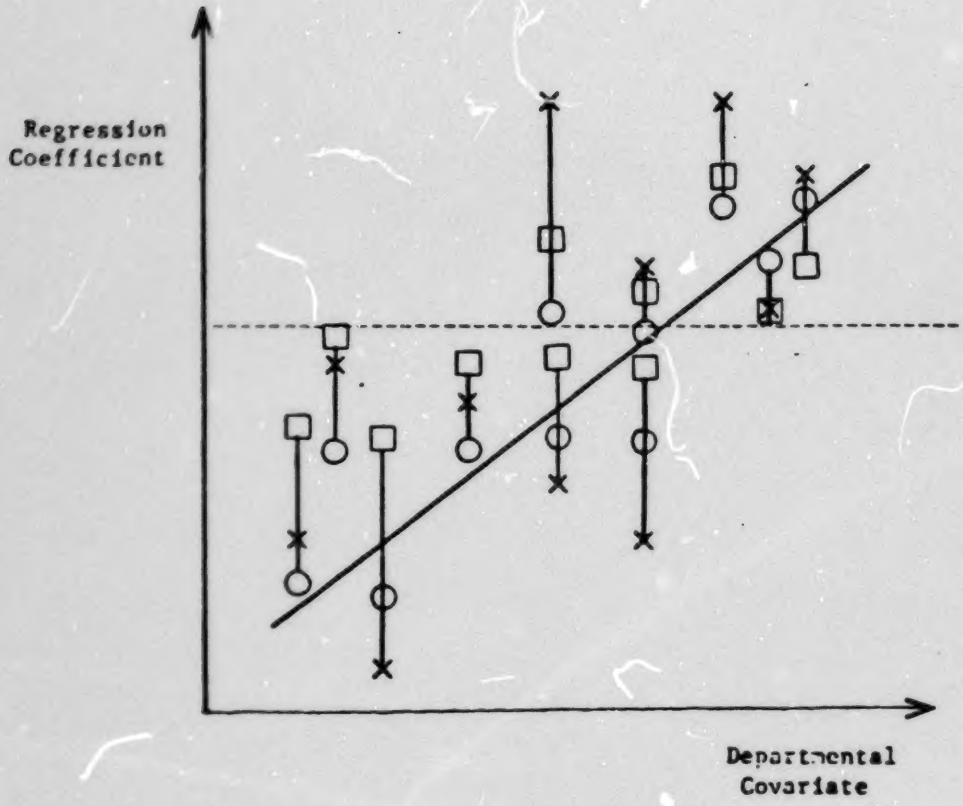
When the model (1), (3), and (4) is employed, the higher order parameters G and $\Sigma^*$ must be estimated. The least squares estimate $\hat{B}_i$ is pulled toward the point $Z_iG$ on the line ZG to yield the empirical Bayes estimate $\tilde{B}_i$.

Figure 1 displays the difference between the consequences of assuming (1) and (2) or (1), (3) and (4). Under (1) and (2), the $\hat{B}_i$ values are pulled toward a common point while under (1), (3) and (4) they are pulled toward the line.

Of course, (3) can be modified to include more covariate information (that is a vector of departmental factors), as well as a vector of Bs rather than a single coefficient. As the heterogeneity of the departments within a proposed cluster increases, presumably more relevant covariate information must be captured in order to preserve the applicability of the empirical Bayes methodology.

From the point of view of validity assessment, the necessity of constructing clusters of departments and carrying out separate analyses for each is something of a nuisance. One of the goals of the present study is to determine the extent to which one can dispense with clusters by incorporating sufficient covariate

Figure 1: Effects of Empirical Bayes Estimation
(Illustrative)




× : Least Squares Estimate

□ : Empirical Bayes Estimate,
Shrinking to a Point

○ : Empirical Bayes Estimate,
Shrinking to a Line

information into the model through assumption (3). A clear limitation is the amount of such information available. In this setting, we are constrained essentially to covariates derived from the predictor scores of students in the department. In other applications, covariates in (3) need not bear any relation to the predictors appearing in (1).

3.3 Descriptions of Models

We will be concerned with the estimation of prediction equations for individual graduate departments. For department i, these equations take the form

$$(5) \qquad Y_{ij} = B_{io} + B_{i1} V_{ij} + B_{i2} Q_{ij} + B_{i3} U_{ij} + e_{ij}$$

where j indexes students within departments. The three predictors are GRE verbal and quantitative scores and UGPA. In what follows, the verbal and quantitative scores have been rescaled by dividing by 200 so that their regression coefficients should be of comparable magnitude to that of UGPA, which is on a 0-4 scale. As usual, the random errors are assumed to be independently normally distributed, and interest centers on the estimation of the vector of parameters:

$$B_i = (B_{io}, B_{i1}, B_{i2}, B_{i3})'.$$

The method in current use, least squares based on data from a single department, will be the standard against which the new procedures will be judged. The various empirical Bayes procedures we propose will differ along two basic dimensions: (1) whether the full family of departments or only clusters of departments are treated simultaneously, and (2) the geometrical structure of the shrinking, that is, to a point, plane, or quadratic surface. In

discussing the first point, we shall generally employ the same set of five clusters of departments: humanities, social sciences (excluding psychology), psychology, biological sciences, and physical sciences. Table 1 presents the total number of departments belonging to each cluster and participating in the Validity Study Service, as well as the total number of students in each cluster.

To enhance readability, we will employ the mnemonic EBxx to designate the various empirical Bayes models. The third symbol will be F or C depending on whether the full family or clusters of departments are used. The fourth symbol will be p,f,q, or q', depending on whether the least squares estimates are pulled toward a point, a plane (flat surface), or one of two quadratic surfaces. Thus, for example, EBFf denotes the approach involving the full family of departments and shrinking toward a plane.

While the clustering feature of these models is easy to describe, the geometrical nature of the shrinking is more difficult. The general structure takes the form

$$B_i' = Z_i'G + D_i, \tag{6}$$

where

$$D_i \sim N(0, \Sigma^*). \tag{7}$$

It remains to specify the covariates included in $Z_i$, which in turn determine the dimension of G, the matrix of unknown coefficients to be estimated. Note that given the level of clustering, equations (5), (6), and (7) determine the model.

TABLE 1

Numbers of Departments and Numbers of Students by Cluster: Total and Half-Samples

| | Total | | Half-Sample 1 | | Half-Sample 2 | |
|---|---|---|---|---|---|---|
| | # Depts. | # Students | # Depts. | # Students | # Depts. | # Students |
| | 190 | 6,946 | 142 | 3,172 | 142 | 3,168 |
| Biological Sciences | 25 | 916 | 23 | 445 | 23 | 440 |
| Humanities | 27 | 753 | 19 | 330 | 19 | 335 |
| Physical Sciences | 46 | 1,379 | 25 | 548 | 25 | 550 |
| Psychology | 23 | 888 | 21 | 434 | 21 | 429 |
| Social Sciences | 69 | 3,010 | 54 | 1,415 | 54 | 1,414 |

We now present the different choices for $Z_i$.

EBCp, EBFp: $Z_i$ is a constant. Thus G is identically one and the model reduces to the form

$B_i \sim N(\mu^*, \Sigma^*)$.

Recall that for EBCp, $\Sigma^*$ is estimated separately for each cluster.

EBCf, EBFf: $Z_i$ is a four component vector of departmental covariates: a constant ($Z_{io}$), mean verbal score ($Z_{i1}$), mean quantitative score ($Z_{i2}$), and mean UGPA ($Z_{i3}$). (Means are taken over all students in the department.) Correspondingly, G is a 4x4 matrix of unknown parameters and $\Sigma^*$ is a 4x4 covariance matrix that is also unknown. For EBCf, G is estimated separately for each cluster but, for reasons of parsimony, a common estimate of $\Sigma^*$ across clusters is employed.

EBFq: $Z_i$ is a seven component vector including $Z_{io}$, $Z_{i1}$, $Z_{i2}$, $Z_{i3}$, $Z_{i1}^2$, $Z_{i2}^2$, $Z_{i3}^2$. Such a model allows us to explore the possibility of a nonlinear relation between the regression coefficients of interest and the departmental covariates. The corresponding EBCq model is not considered because of the number of coefficients to be estimated.

EBFq': $Z_i$ in EBFq is augmented by three interaction terms $Z_{i1} Z_{i2}$, $Z_{i1} Z_{i3}$, $Z_{i2} Z_{i3}$.

It should be noted that EBCp is really a special case of EBCf with a single covariate, the constant. In its full generality,

assumption (6) implies that the vectors of coefficients in the prediction equation (5) are not normally distributed about a point in 4-space, but rather are normally distributed about a plane in 8-space with the typical (average) set of coefficients being linearly related to the level of achievement of the average student in the department, as measured by GRE verbal and quantitative scores, and UGPA. The nature of this linear relationship is determined by the value of G.

How closely the apparent values of $B_i$ fall to the plane Z'G is roughly indicated by the size of the diagonal elements of $\Sigma^*$. All other things being equal, a smaller estimated $\Sigma^*$ results in the least squares estimates $\hat{B}_i$ being pulled more strongly toward the plane.

We did not expect EBCp to work well, given the heterogeneity of departmental types within four of the clusters. Rather, it was introduced as a benchmark against which the performance of the other empirical Bayes procedures could be measured. On the other hand, EBCf seemed to possess sufficient flexibility to offer some promise of reasonable performance.

Models EBFq and EBFq' postulate that the vectors of coefficients $B_i$ are generated from a quadratic surface rather than a plane. That is, these models permit a nonlinear relation between the coefficients and the covariates. Our plan has been to study the performance characteristics of EBFq' and, if they proved favorable, to investigate the possibility of progressively simplifying it, that is, employ EBFq, EBFf, or even EBFp. From the point of parsimony, we would prefer to use a model from the EBF family provided its performance essentially

matches that of the EBC family. In Section 3.5, we will discuss various measures of performance.

Before turning to that section, we will describe some features of the parameter estimates. It should be noted that we employ maximum likelihood estimation, using the EM algorithm (Dempster, Laird, & Rubin, 1977) as a computational technique for obtaining estimates of the hyperparameters G and $\Sigma^*$. As a by-product, the posterior distributions of $B_i$, given the data and the maximum likelihood estimates, G and $\Sigma^*$ are produced. The means of these posterior distributions serve as the empirical Bayes estimates of $B_i$. The essential details have already been published (Braun, Jones, Rubin, & Thayer, 1983), while the Appendix contains additional material relevant to the present application.

3.4 Model Description

One of the drawbacks of the more complex empirical Bayes models we have introduced is that they are difficult to visualize and understand. Even the relatively simple EBFf requires a fair amount of effort before we can draw some insight from its numerical characteristics. To illustrate, in this section we will carry out an analysis of EBFf and compare it to least squares.

Recall that for EBFf we assume that the vectors of regression coefficients for the departments are themselves generated from a linear regression on the mean test scores of the students in the department. This multivariate linear regression is characterized by a matrix of coefficients, denoted G, and a normal error structure characterized by a covariance matrix, denoted $\Sigma^*$. Table 2 presents

**TABLE 2**

**Numerical Characteristics of a Fitted Empirical Bayes Model (EB2b)**

$\hat{G}$ : Estimate of Matrix of Regression Coefficients at Hyperparameter Level

$$\begin{bmatrix} 1.60 & 1.20 & -.03 & -.29 \\ -.66 & .24 & .08 & -.15 \\ -.70 & -.25 & .34 & -.01 \\ -.53 & -.26 & -.25 & .41 \end{bmatrix}$$

$\hat{\Sigma}^*$: Estimate of Covariance Matrix at Hyperparameter Level

$$\hat{\Sigma}^* = 10^{-2} \times \begin{bmatrix} 9.74 & 1.01 & .07 & -3.85 \\ & .60 & .04 & -.83 \\ & & .17 & -.20 \\ & & & 2.03 \end{bmatrix}$$

$\hat{\Sigma}^*$ (correlation form)

$$\begin{bmatrix} 1 & .42 & .05 & -.86 \\ & 1 & .13 & -.75 \\ & & 1 & -.34 \\ & & & 1 \end{bmatrix}$$

the estimates $\hat{G}$ and $\hat{\Sigma}^*$ of these parameters for the GRE Validity Study Service data described in Section 2.

One interpretation of G is that, on the average, a department with specific mean test scores $\bar{v}$, $\bar{q}$, and $\bar{u}$ should have, according to the fitted model, the following regression coefficients for the constant, GRE-V, GRE-Q, and UGPA:

$$B^*_{io} = 1.60 - .66\,\bar{v} - .70\,\bar{q} - .53\,\bar{u},$$

$$B^*_{i1} = 1.20 + .24\,\bar{v} - .25\,\bar{q} - .26\,\bar{u},$$

$$B^*_{i2} = -.03 + .08\,\bar{v} + .34\,\bar{q} - .25\,\bar{u},$$

$$B^*_{i3} = -.29 - .15\,\bar{v} - .01\,\bar{q} + .41\,\bar{u}.$$

Of course, actual departments with these mean test scores will have true regression coefficients that differ from those above because of the variation between departments, expressed by the error component D in equation (3). Empirical Bayes tries to estimate the true coefficients by combining the information in the least squares estimates and the "typical estimates" $B^*_{io}$, $B^*_{i1}$, $B^*_{i2}$, $B^*_{i3}$.

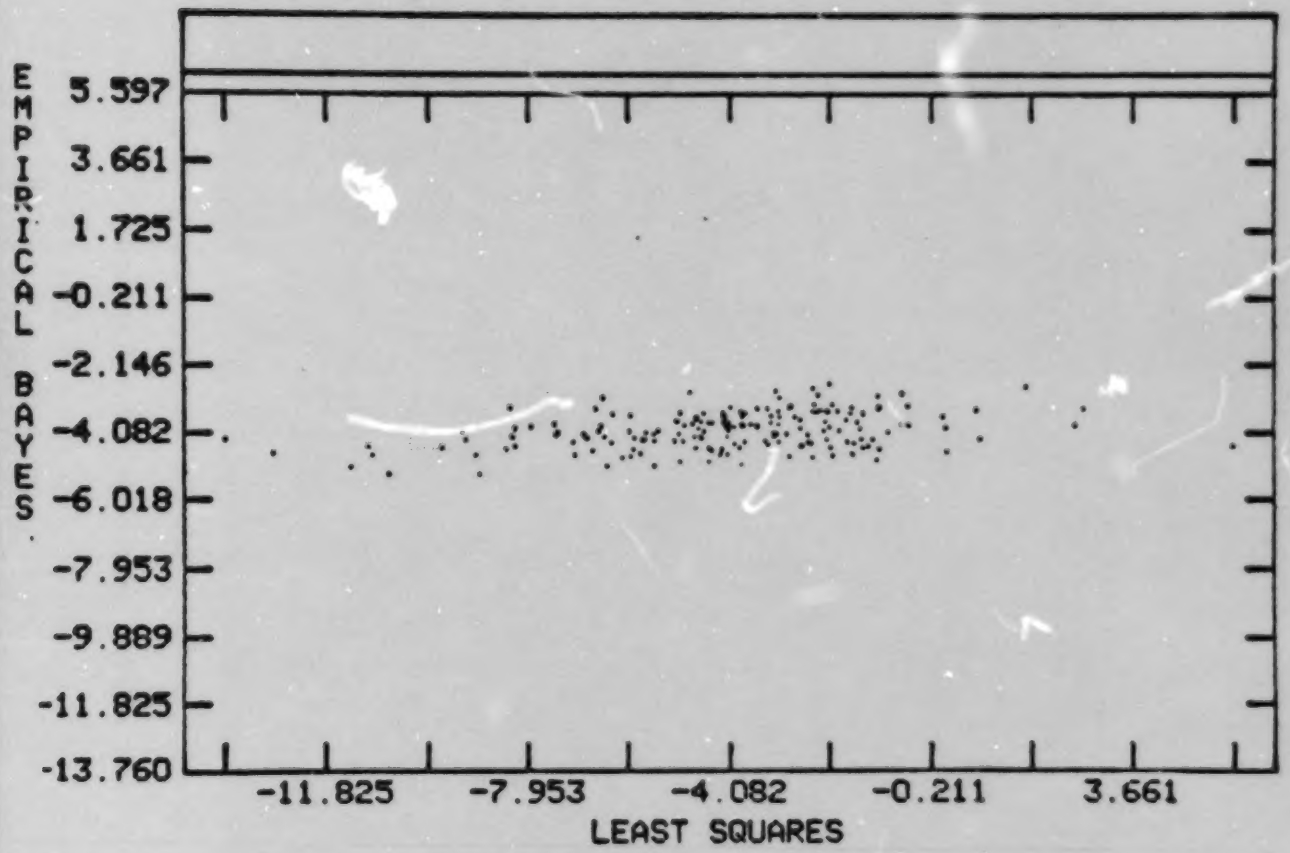
Interestingly, the diagonal elements of the estimate of $\Sigma^*$ are rather small, compared to the estimated variance of the least squares estimates. Accordingly, the empirical Bayes estimate $\gamma_{B_i}$ falls nearer to the typical estimate $B^*_i$ than to the least squares estimate of $B_i$. In other words, the apparent variability in the least squares estimate is so large that the empirical Bayes compromise leads to shrinking the least squares estimate almost all the way to the plane. This is illustrated schematically in one

dimension in Figure 1. Each department's estimated prediction equation under empirical Bayes remains unique because each department possesses a unique set of mean test scores. However, two departments with the same mean test scores would obtain estimated prediction equations from empirical Bayes that would be nearly the same. That is, the differences between the least squares prediction equations for the two departments are largely attributed to random variation and are eliminated by the empirical Bayes process. It should be emphasized that these findings are particular to this study; in other applications, the precision of the individual unit's least squares estimate may be comparably greater and, as a consequence, have a greater direct influence on the corresponding empirical Bayes estimate.

Inspection of $\hat{G}$ reveals that, in general, we should expect that the coefficient in the departmental prediction equation associated with a particular test score will increase as that department's mean test score increases, that is, the diagonal elements of $\hat{G}$ are all positive. Furthermore, that same coefficient will decrease as the other mean test scores increase, though this effect tends to be smaller. Since the intercept decreases with increasing mean test score, we may say roughly that, as the mean test scores increase across a series of departments, the corresponding prediction planes become somewhat steeper and, as there appears to be no association between the size of the slope and the estimate of residual variance, the quality of the fit becomes somewhat better.

Figure 2

Scatterplot of Intercept Coefficients for Two Models: All Departments



The diagonal elements of $\hat{\Sigma}^*$ clearly show that across departments, $B_{io}$ displays the most dispersion, followed by $B_{i3}$, $B_{i1}$, and $B_{i2}$ in that order. The correlation matrix derived from $\hat{\Sigma}^*$ also provides useful information. We note that the correlations among the intercept, the verbal and quantitative coefficients are all positive and modest in size compared to the rather large negative correlations they exhibit with the UGPA coefficient. Although we have not developed a formal model to explain this pattern, it is very likely an outgrowth of the selection processes operating at the departmental level. For example, such a pattern would likely develop if admissions were based more heavily on the GRE or UGPA, but not both, in most departments.

To develop an appreciation for the effect of the empirical Bayes approach, we display in Figure 2 a scatterplot of the least squares against the empirical Bayes (EBFf) estimates of the intercepts of the within-department prediction planes. Although the two sets of estimates are strongly associated, the dispersion of the former is considerably greater than that of the latter, the ratio of the standard deviations of the marginal distributions being about five to one. These aspects of the empirical Bayes models are poor substitutes for more informative ways of looking at eight-dimensional space. Unfortunately, such views are difficult to capture on two-dimensional surfaces.

3.5 Model Selection

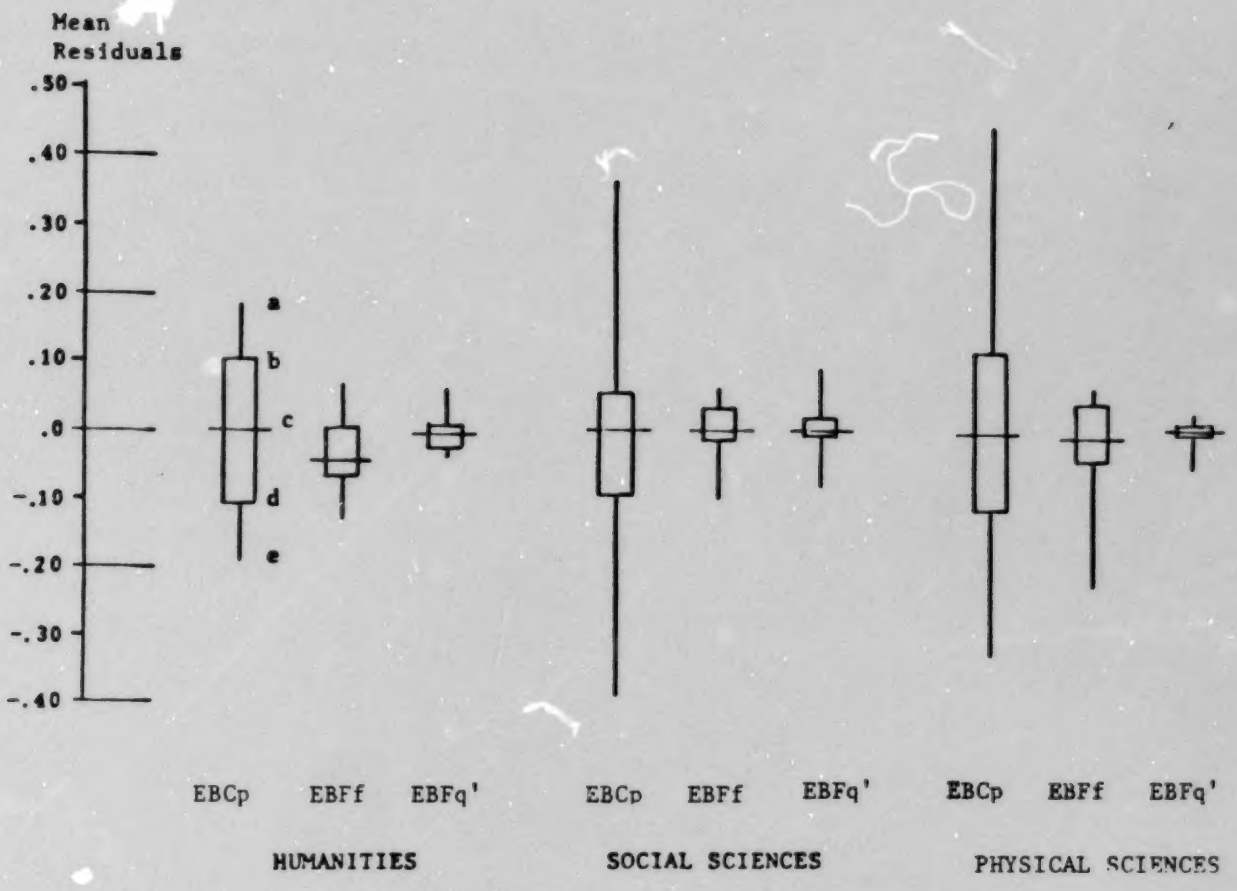
The selection of one among many complex models in this setting involves a number of criteria related to both goodness-of-fit and stability, as well as nonstatistical considerations. Our approach has been to screen the available empirical Bayes models with a

view to winnowing out all but two or three, and then to compare the performance of those remaining with that of least squares on the basis of cross-validation studies. We shall present our analysis in approximate historical sequence.

Because empirical Bayes estimates for a given department are not just based on data from that department alone, they do not appear to fit the data as well as least squares estimates based solely on the department's data. For example, least squares estimates have the property that the mean residual over all students in the department must be zero. That is, the average predicted FYA and the average observed FYA are equal. In general, empirical Bayes estimates do not share this property. Thus, one dimension along which we can compare different empirical Bayes models is how close they come to the ideal of producing zero-mean residuals.

Three models were chosen: EBCp, EBFp, and EBFq'. Although only EBCp makes use of the cluster structure, for each method the mean residuals by department are grouped by cluster for purposes of comparison. The mean residual for a department is the difference between the average observed FYA and the average predicted FYA derived from the method under study. Figure 3 displays box-and-whisker plots (Tukey, 1977) of these sets of mean residuals for three of the clusters. As one would hope, the distribution of mean residuals in each cluster-method combination is approximately centered about zero. However, there are large differences in variability among the distributions. Since less variability is to be preferred, EBCp must clearly be rejected as a viable alternative. Apparently, the heterogeneity

Figure 3: Box-and-Whisker Plots of Mean Residuals for Three Models




NOTE: A box-and-whisker plot displays five important characteristics of a distribution. They are:

a = maximum
b = upper quartile
c = median
d = lower quartile
e = minimum

among departments within a cluster is sufficient to preclude a simple shrinking-to-a-point model from performing adequately.

EBFf performs somewhat more poorly than EBFq' but requires fewer parameters. Since the ultimate goal of the analysis is to produce models that can predict FYA well, we prefer to consider other aspects of the models before deciding to discard one or the other. One such feature is the scatter of the empirical Bayes estimates of the vector of regression coefficients about the surface from which the true vectors were apparently generated. Excessive variability or systematic patterns in the scatter may indicate deficiencies in the formulation of the higher level of the model--equation (6).

We have chosen to consider two statistics. Let

$$g_i = ||\tilde{B}_i - Z_i'\tilde{G}||$$

be the Euclidean distance (in parameter space) between the point representing the empirical Bayes estimate of the true vector of regression coefficients for the department and the corresponding point on the estimate of the surface from which these vectors were generated. Since $g_i$ gives equal weight to each component, we also constructed the statistic

$$h_i = (\tilde{B}_i - Z_i'\tilde{G})' Z_i,$$

which measures the difference in predicted FYA for a student with average test scores, obtained by using $\tilde{B}_i$ or $Z_i'\tilde{G}$. Here the componentwise differences are weighted by the mean scores in the department. Although $g_i$ and $h_i$ may be studied in isolation, we decided to plot them against the quantity

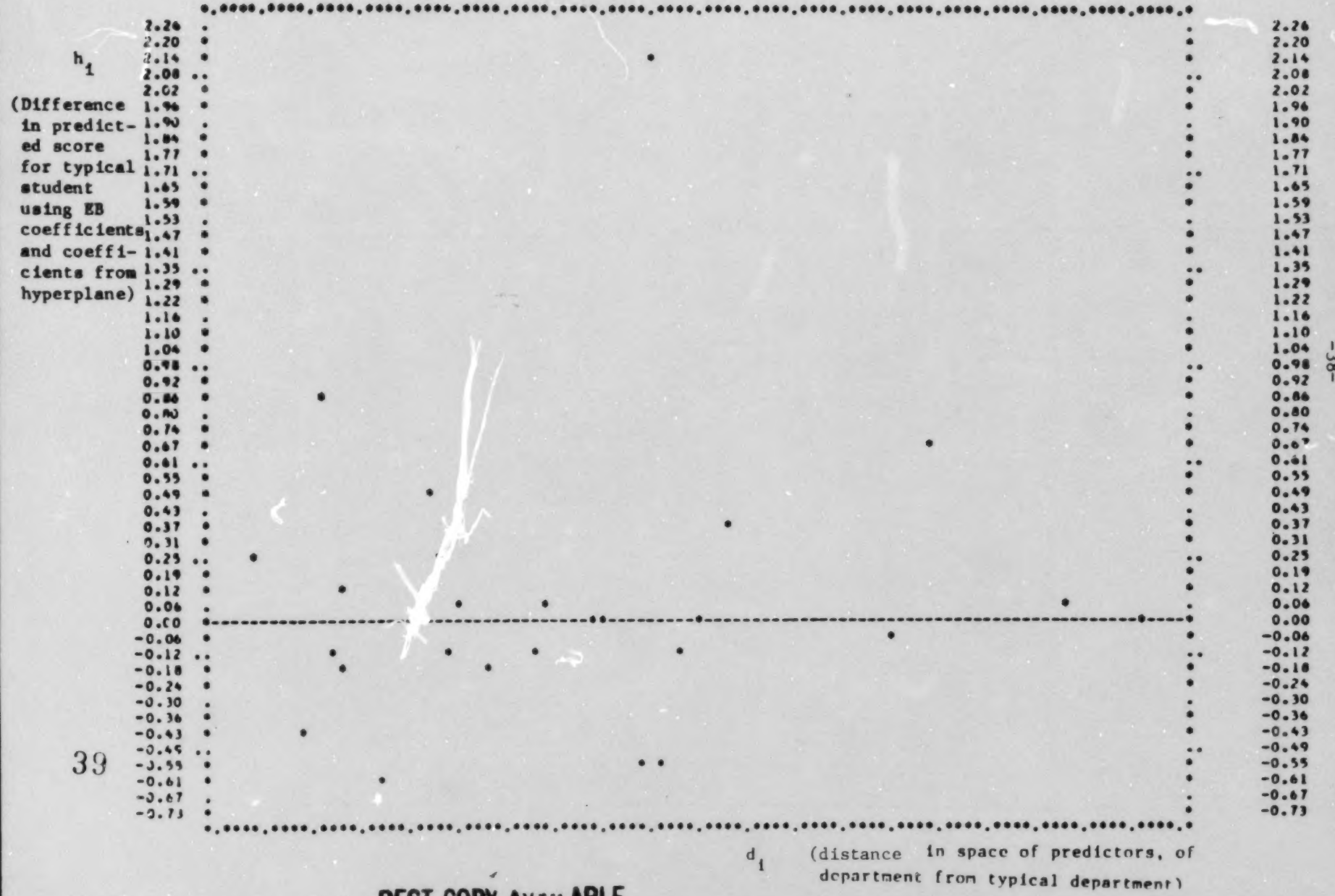
$$d_i = ||w_i - C||$$

where $w_i = (Z_{i1}, Z_{i2}, Z_{i3})'$ is the vector of mean scores in the department on GRE-V, GRE-Q, and UGPA, and C is obtained by taking the median, componentwise, of these means across departments in the cluster. Thus, the first component of C is the median of the average verbal scores for the departments in the cluster. Large values of $d_i$ indicate that department i is atypical, in some sense, of the departments in the cluster.

Plots of $g_i$ or $h_i$ against $d_i$ may indicate whether there are systematic patterns in the quality of fit of the model. For this reason, we construct these plots separately for each cluster, even though the empirical Bayes model we employ may not utilize the cluster structure at all. We could more easily construct a simple plot aggregating over all clusters. But the present approach will allow us to detect difficulties at the cluster level.

Plots of $g_i$ against $d_i$ were generated for three models: EBCf, EBFf, and EBFq'. Interestingly, there were no systematic patterns evident except for a tendency for large values of $d_i$ (atypical departments) to be associated with the small value of $g_i$. Figure 4 provides a representative illustration. On the other hand, the values of $g_i$ for EBFf were considerably larger than for the other two models. EBFq' performed somewhat more poorly than EBCf, but the latter requires many more parameters. Similar comparisons were carried out on the basis of plots of $h_i$ against $d_i$. Differences among methods were somewhat reduced, and again there were no evident systematic patterns. Consequently, no decision to reduce the number of models being considered was made at this point.

Figure 4: EBCf for Humanities Cluster. $h_i$ vs. $d_i$










BEST COPY AVAILABLE

Since EBFq' does involve the estimation of a large number of parameters, we explored the possibility of eliminating some of the covariates appearing in (6). In particular, using slight generalizations of the sensitivity analyses advocated in Belsey, Kuh, and Welsch (1980), we were able to demonstrate the redundancy of the cross-product terms $Z_{i1} Z_{i2}$, $Z_{i1} Z_{i3}$, $Z_{i2} Z_{i3}$. Consequently, we began the series of cross-validation tests with four models: least squares, EBCf, EBFf, EBFq. Among the empirical Bayes models considered, EBFf is the most parsimonious in terms of the number of parameters to be estimated.

3.6 Cross-Validation

Cross-validation as a model selection technique has a long and honorable history (Stone, 1978). Essentially, the idea is to use a model fitted on one set of data to predict the results for an independent set of data. In practice, replicate data sets are rare, and the usual remedy has been to split the existing data set in half, fitting the model on one half and using it to make predictions for the other half. Of course, the drawback is that the estimated coefficients will be more variable than if they were fitted using the entire data. Nonetheless, the method gives a good indication of which models are too sensitive to artifacts in the data at hand.

Our approach has been to split each department in half at random and, thus, to construct two half-samples of the full validity data base at the departmental level. One half-sample is selected to be the calibration sample; least squares and three empirical Bayes methods are applied to it. The resulting fitted models, together with the predictor information for students in the other half-sample, the

cross-validation sample, are used to predict FYAs that are compared to the actual FYAs obtained by those students. Then the roles of the two half-samples are reversed and the process repeated.

In carrying out the construction of the half-samples, we eliminated departments with fewer than 10 students so that each half-sample would have at least five students. As a result, only 142 of the original 190 departments (6,340) students were included in the cross-validation. Table 1 displays the breakdown by cluster.

There are a number of ways to assess success in prediction. In the area of measurement, it is common to correlate the predicted FYAs with the observed FYAs. The higher the correlation, the more useful the predictions are held to be. While this may be reasonable in some settings, it is also true that some systems that produce very large errors of prediction may nevertheless yield high correlations with the observed scores. It should also be noted that in the cross-validation setting, the familiar relation between the square of the correlation and the proportion of variance explained no longer holds. Thus, we prefer quantities related to the mean squared error of prediction as our criteria for determining the usefulness of a model.

We have employed two related measures of agreement. The first, root mean square deviation (RMSD), is commonly used by statisticians. Suppose there are m students in a half-sample of the department. Let $y$ denote a generic obtained FYA and $\tilde{y}$ a generic estimated FYA based on coefficients from the other half-sample as well as the predictor scores of the student. Then,

$$\text{RMSD} = [m^{-1} \sum_{j=1}^{m} (y_j - \tilde{y}_j)^2]^{1/2}.$$

Like the familiar standard deviation, it assesses the prediction error on the same scale the data are measured on.

A second measure is the difference between predicted and actual $R^2$, which we will denote by $DR^2$. In regression analysis, $R^2$ is the proportion of variance in the criterion that is explained by the regression. Given a fitted model, one can calculate what the $R^2$ would be for a new data set following the same model but with a different set of predictor scores. We call this the predicted $R^2$. More precisely, we obtain the predicted $R^2$ for a regression derived through some procedure (least squares, empirical Bayes, and the like) applied to the calibration sample by assuming that the least squares fit to the cross-validation sample will yield exactly the same characteristics (slopes and residual variance). Combining those characteristics with the distribution of predictor scores in the cross-validation sample yields a value of $R^2$ for the putative least squares fit. We are acting as if the fitted regression to the calibration data will provide a good estimate of the least squares fit to the cross-validation data. This places, perhaps, an extra burden on the empirical Bayes fits. With the new data in hand, the variation in the residuals, $y-\tilde{y}$, can be compared to the variation in y. The actual $R^2$ is

$$1 - \frac{RMSD^2}{\text{variance (y)}}$$

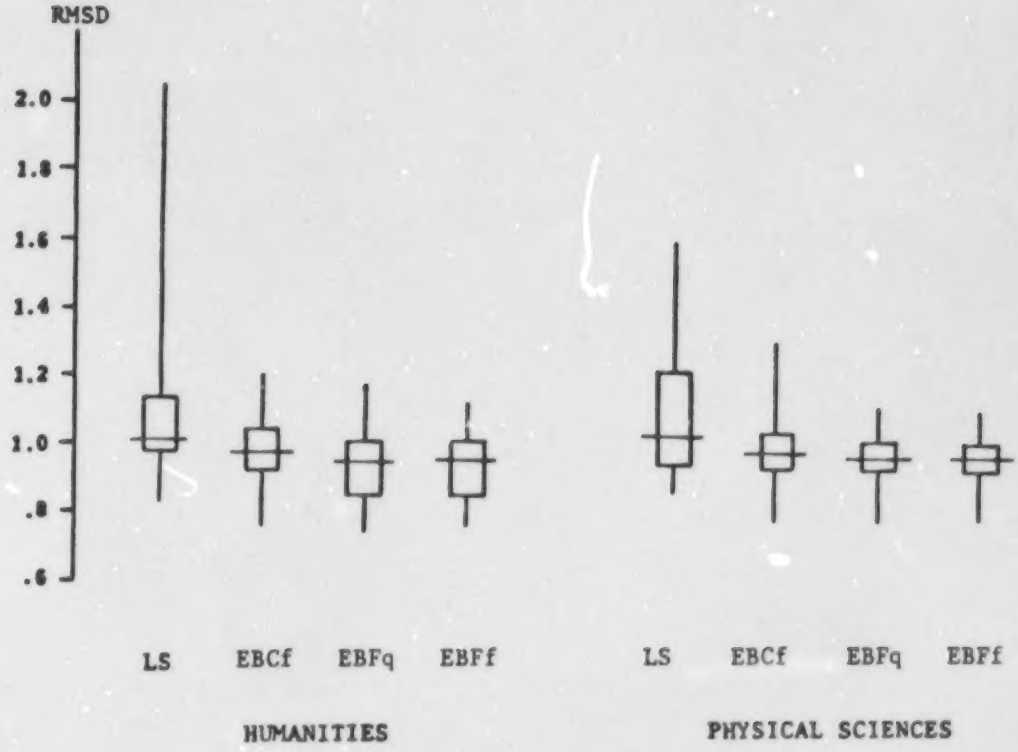
This is something of a misnomer since the actual $R^2$ compares the size of the mean squared error of prediction to the variance of the criterion. It will be negative when the predictions $\tilde{y}$ are inferior to the mean. The $DR^2$ is defined as

$$DR^2 = \text{predicted } R^2 - \text{actual } R^2.$$

A positive value of $DR^2$ indicates that the actual performance of the fitted model is not as good as would have been expected from the original data. One would expect such behavior of estimates based on least squares because of the potentially great effect that idiosyncrasies of the particular sample may have on the least squares estimates. On the other hand, superior procedures should exhibit $DR^2$ values approximately symmetrically distributed about zero with only modest dispersion. That is, the predicted performance is a reliable estimate of the actual performance. In a sense, RMSD measures how accurate the predictions are, while $DR^2$ measures how well we can guess, a priori, at the quality of those predictions.

Because of the volume of information generated in this segment of the analysis, only a small representative sample of the findings can be presented. As before, we display the results grouped by cluster so that particular patterns can be discerned. In all five clusters, least squares estimates perform more poorly on both measures of accuracy of prediction than all three empirical Bayes estimates. On RMSD, least squares is typically about 10 percent larger than the empirical Bayes methods, which are nearly equivalent. Figure 5 contains box-and-whisker plots that schematically display key features of the distributions of RMSD values for the four methods for two clusters, humanities and physical sciences. Note that in each cluster, the maximum RMSD for least squares is exceedingly large and certainly dominates the maximum exhibited by the empirical Bayes methods.

Figure 5: Box-and-Whisker Plots of Root Mean Square Deviations (RMSD) for Four Models. Cross-Validation Analysis

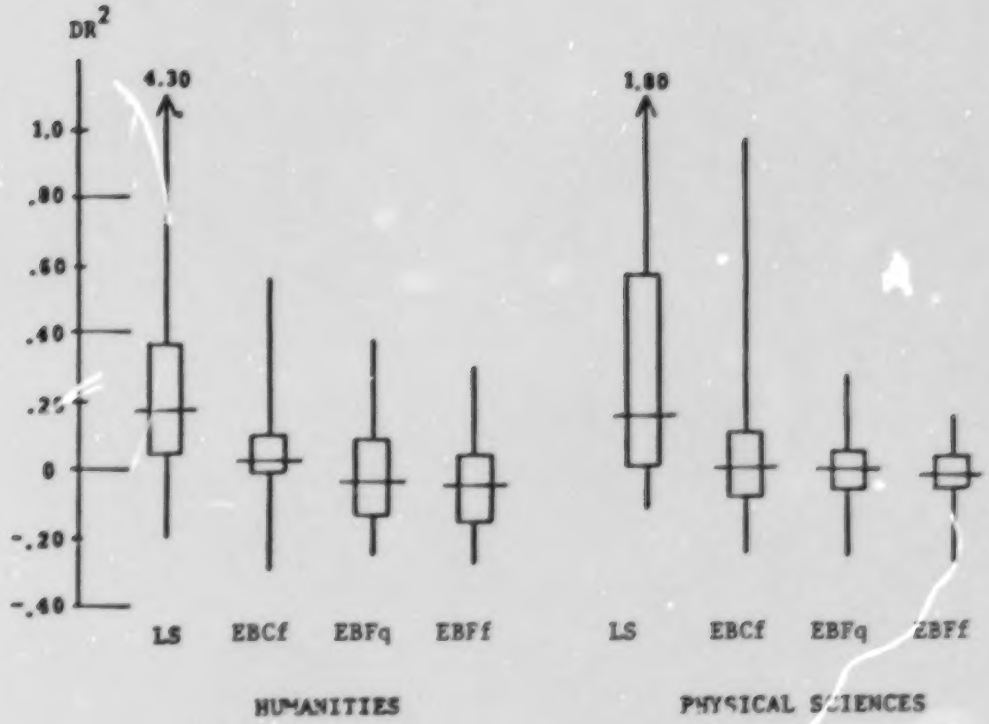


Because the RMSD is measured in units of FYA, its magnitude can be assessed for practical significance. The typical value of RMSD seems to be about one standard deviation, corresponding to a fair amount of difference between predicted and observed. The largest values, about two standard deviations, are indicative of essentially useless predictions.

As we have noted above, EBCf, EBFf, and EBFq are nearly equivalent, with EBCf perhaps being somewhat worse. Since EBFf is the most parsimonious in terms of the number of parameters to be estimated, it is preferred.

Consideration of $DR^2$ leads even more unambiguously to the same conclusion. Figure 6 presents box-and-whisker plots for the humanities and physical sciences clusters of the distributions of $DR^2$ for the four methods. The inferiority of least squares is apparent. The predicted $R^2$ tends to grossly overestimate the actual $R^2$, resulting in large positive values of $DR^2$. In fact, more than three-quarters of the values are positive. By contrast, the distribution of $DR^2$ for the empirical Bayes methods is more nearly symmetric about zero although EBCf does have a tendency to produce some modestly large values. The superiority of EBFf is evident. Given our definition of predicted $R^2$, this result is slightly surprising. Nevertheless, its clear implication is that empirical Bayes fits were more reproducible than least squares fits. Simply put, the predicted FYAs from empirical Bayes were closer to the mark than those of least squares, and the quality of the fit could itself be more accurately predicted for empirical Bayes.

Figure 6: Box-and-Whisker Plots of $DR^2$ for Four Models. Cross-Validation Analysis



3.7. Supplementary Analyses

Of course, we have by no means exhausted the family of prediction systems with which empirical Bayes can be compared. For example, it has been suggested (Wilson, 1982) that the data from all departments in a given discipline be pooled and a single least squares regression plane be fitted. We have carried out some exploratory analyses in this direction by considering four disciplines: psychology, economics, chemistry, and physics. Using the same cross-validation scheme described above, we compared the predictions of EBFf with those of the pooled least squares plane for each of the four disciplines. In each case EBFf proved superior. Another drawback to this kind of pooling is that many disciplines are represented by only a few departments in the Validity Study Service data so that further pooling across disciplines is required.

Another approach consists of grouping departments by the characteristics of their students. One suggestion (Burton, 1982) is to use the difference between mean GRE verbal and mean GRE quantitative scores. A simple scheme involves the formation of two clusters, according to whether this difference is negative or positive. The data within each cluster is pooled and a single least squares regression plane fitted for each cluster. Again, empirical Bayes performed better overall in cross-validation.

Finally, we investigated the possibility of including among the departmental covariates other features of the distribution of predictor scores of the students such as variances and covariances. An important motivation for this step was the concern that apparent differences

between departments in the magnitude of the coefficients in the prediction equations may be due to variation in the amount of restriction of range experienced. However, the inclusion of predictor score variances did not result in any improvements. Thus, it does not appear that our conclusions have been driven by differential restriction of range.

## 4. Subgroup Analyses

### 4.1 Introduction

One question of considerable interest concerns the possible differential efficacy of the GRE battery or UGPA as predictors of FYA across various subgroups of the population. These subgroups may be defined by sex, race, age, mother tongue, or some combination of these factors. Classification of students by these factors, however, further exacerbates the problem of small sample sizes. In fact, it is usually impossible to obtain separate prediction equations for each group in each department by ordinary least squares methods.

Bayesian methods, and empirical Bayes methods in particular, provide a reasonable solution to this problem. In the context of the Graduate Management Admission Test, Braun and Jones (1981) demonstrated that separate prediction equations for White students and Black students could be estimated in each institution, even though Blacks were less than 6 percent of the cohort. Technically, the extension of the model of Section 3 to encompass this application is very simple.

Suppose, for example, that students are categorized according to two factors, each at two levels. For definiteness, suppose these

factors are sex (male or female) and age (less than 25, more than 25). We then assume that

$$(8) \quad Y_{ij} = B_{io} + B_{i1} V_{ij} + B_{i2} Q_{ij} + B_{i3} U_{ij} + I^1_{ij} \left\{ B_{i4} + B_{i5} V_{ij} + B_{i6} Q_{ij} + B_{i7} U_{ij} \right\} + I^2_{ij} \left\{ B_{i8} + B_{i9} V_{ij} + B_{i10} Q_{ij} + B_{i11} U_{ij} \right\} + e_{ij}$$

where

$$I^1_{ij} = \begin{cases} 1, \text{ if student is female} \\ 0, \text{ if student is male} \end{cases}$$

and

$$I^2_{ij} = \begin{cases} 1, \text{ if student is more than 25} \\ 0, \text{ if student is less than 25} \end{cases}$$

As before, i indexes departments and j indexes students within departments. The higher level of the model remains the same. The expanded vector of coefficients $B_i = (B_{io}\ B_{i1} \ldots B_{i11})'$ is assumed to follow a regression of the form $B_i = Z_i'G + D$. Equation (8) allows us to fit a different prediction plane for each sex-age combination, the only restriction being that any one of the planes can be linearly determined from the other three. (It would require three indicator functions to fit four independent planes. Given the large number of coefficients to be estimated, this extension was not followed through.)

The indicator functions $I^1$ and $I^2$ determine which set of coefficients comes into play. For example, for males over 25, $I^1 = 0$ and $I^2 = 1$ so the plane is determined by $B_{io}, \ldots B_{i3}$ and by $B_{i9}, \ldots B_{i11}$. For females over 25, $I^1 = 1$ and $I^2 = 1$ so the plane is determined by all twelve coefficients.

Another way to think of the process is to imagine that the first four coefficients determine the basic prediction plane for males under 25. The other two sets of four coefficients represent the modifications that must be made to the basic plane to account for differences for females or older students, or both.

A more conservative approach is to fit a model in which the separate planes are constrained to be parallel. Such a model would take the form

$$(9) \qquad Y_{ij} = B_{io} + B_{i1} V_{ij} + B_{i2} Q_{ij} + B_{i3} U_{ij} + I^1_{ij} B_{i4} + I^2_{ij} B_{i5} + e_{ij}.$$

The inclination of the plane remains the same for all groups, and only its height may vary. A more complete treatment of fitting empirical Bayes models to sparse data may be found in Braun et al. (1983).

Model EBFq was employed as the basis of our studies in this section. Thus the higher level of the model may be represented by a quadratic surface determined by linear and pure quadratic terms. When the departmental regression takes the form (9), we denote the model by EBFqe; when it takes the form (8), we denote the model by EBFqe'.

4.2 Results

To examine the viability of differential prediction systems, we chose to group individuals by age and sex as described above. At first, to eliminate confounding effects we selected only individuals who were White and for whom English was the primary language of communication. Unfortunately, this reduced the total sample by about 60 percent. We, therefore, abandoned this selection and employed all students with complete information on FYA, GRE verbal and quantitative

scores UGPA, age, and sex. The total number of students was 5,491, somewhat lower than the original 6,946. Table 3 displays the breakdown by sex and age for each half-sample. Again, departments with fewer than 10 students were excluded from the final cross-validations.

Half-samples were generated as before, with the eligible students in each department being divided into two equal groups. The models to be compared, least squares, EBFq, EBFqe, and EBFqe', were each fitted to one half of the data base and then applied to the other half. Only the RMSD measure was calculated. In general, the performance of the empirical Bayes models was superior to that of least squares, but EBFq, which does not take account of age or sex, did better than EBFqe and EBFqe', which do. Thus, on the basis of the cross-validation, there is no reason to employ separate prediction planes for the different age-sex groups.

On the other hand, there appears to be a consistent pattern in the fitted coefficients that may merit further investigation. Under model EBFqe, for a given set of predictor scores, the predicted FYA for females tended to be higher than that for males, while the predicted FYA for those over 25 tended to be higher than that for younger students. The former effect was much more pronounced than the latter. As noted above, neither was borne out in the cross-validation, so we must conclude that a sex effect, if it exists, must be estimated by using more delicate methods.

The analysis was continued by considering differential prediction equations by race. Unfortunately, race was known for only about 55 percent of the original sample. We employed three categories:

TABLE 3

Counts of Students with Complete Information on GRE-V, GRE-Q, UGPA, FYA, Age, and Sex

| | Half-Sample 1 | | Half-Sample 2 | |
|---|---|---|---|---|
| | # Depts. | # Students | # Depts. | # Students |
| Total | 121 | 2,747 | 121 | 2,744 |
| Biological Sciences | 19 | 347 | 19 | 341 |
| Humanities | 16 | 280 | 16 | 283 |
| Physical Sciences | 22 | 505 | 22 | 506 |
| Psychology | 20 | 390 | 20 | 387 |
| Social Sciences | 40 | 1,225 | 40 | 1,227 |

| | Males | Females | Total | Males | Females | Total |
|---|---|---|---|---|---|---|
| Age < 25 | 1,253 | 994 | 2,247 | 1,199 | 1,054 | 2,253 |
| Age > 25 | 18 | 313 | 500 | 201 | 290 | 491 |
| Total | 1,440 | 1,307 | 2,747 | 1,400 | 1,344 | 2,744 |

Whites, Asians and Oriental Americans, and other minorities (principally Blacks). These constituted approximately 88 percent, 2 percent, and 10 percent, respectively, of the reduced sample.

The half-samples were constructed as before and the cross-validation analyses showed again that empirical Bayes outperformed least squares, with EBFq proving superior to EBFqe and EBFqe'. Thus, there appears to be no justification for employing different prediction planes. Consideration of the coefficients in EBFqe indicates a tendency for the predicted FYAs of Asian-American students to be somewhat higher than those of the other students with the same predictor scores.

Our analysis has been carried out in the context of the empirical Bayes formulation. It is quite conceivable that another prediction system, involving more pooling of data, might yield other conclusions, particularly with regard to the question of different prediction equations by race. In this setting, the small sample sizes for minority groups perhaps require that further constraints be placed on the fitted models. More work in this direction should be carried out.

## 5. Clustering

### 5.1 Introduction

As we noted in Section 1, the formation of clusters of departments has always played an important role in the study of the validity of various predictors of graduate school performance. Such clusters are often created by grouping departments by discipline or by sets of related disciplines. Another approach has been to group departments by their location along a verbal-quantitative axis (Wilson, 1979). As we have seen in Section 3, use of five broad clusters of departments did not improve the performance of the empirical Bayes methods. It could be argued that, despite the use of extended models that can accommodate heterogeneity, the five clusters are too diverse and that perhaps 10 or even 15 clusters would be more appropriate. Thus, one would be fitting EBCf, which would require a separate plane--cf. (3) in Section 3.1--to be estimated for each cluster. However, with n clusters, the number of regression parameters to be fitted in equation (3) is n times the number required by EBFf. Since n = 5 did not prove sufficient, it is difficult to believe, on grounds of stability, that higher values of n would be beneficial.

Consequently, we approached the problem from a different perspective. We had information available to us (see Section 2) on the characteristics of the applicants to all departments in the United States for a given year, summarized by department type (discipline). Our aim was to group disciplines by the similarity of the characteristics of their applicants. (Unfortunately, corresponding data for admitted students or enrolled students were not available.) We employed

two different sets of data. The first involves the distribution of types of GRE Subject Tests taken by the applicants and the scores they achieved. Thus, with each discipline is associated a two-factor contingency table containing counts of applicants sorted by test and test score. An example is given in Table 4.

Although there are published methods for carrying out hierarchical clustering of nominal data (Hartigan, 1975; BMD-P-77), we were not satisfied with their properties and devised our own method (Braun & Jones, 1982). The essential difference in our method is that, as we form clusters, we do not pool the data over the disciplines in the cluster but retain all the information for each member. A likelihood ratio statistic is employed to determine which disciplines form the most homogeneous groups with respect to the distribution of the counts in the associated contingency table.

The second set of data we used, which comes from the same source, provides the scores on GRE verbal and quantitative for all applicants to the departments in a discipline. Unfortunately, only the marginal distributions and not the joint distributions were available. That is, for example, we knew how many had verbal scores between 500 and 600 and how many had quantitative scores between 600 and 700, but not how many fell into both categories simultaneously. However, we were able to estimate this joint distribution by assuming that the correlation between verbal and quantitative scores in each discipline equalled the correlation in the entire applicant pool, a known quantity. With each discipline we were able to associate a new two-factor contingency table containing approximate counts of applicants sorted

TABLE 4

Distribution of Students by Advanced Test Taken and Test Score. Genetics

| Test Name | Distribution of Scores | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 200-290 | 300-390 | 400-490 | 500-590 | 600-690 | 700-790 | 800-890 | 900-990 | Total No. of Scores |
| ADVANCED TESTS | | | | | | | | | |
| BIOLOGY | | 6 | 32 | 114 | 239 | [illegible]3 | 69 | 3 | 706 |
| CHEMISTRY | | | | 4 | 6 | 1 | | | 11 |
| EDUCATION | 1 | | 1 | 1 | 1 | | | | 4 |
| ENGINEERING | | | 2 | | | | | | 2 |
| FRENCH | | 1 | 1 | 1 | | | | | 3 |
| GEOLOGY | | | | 1 | | | | | 1 |
| LITERATURE | | | 1 | | | | | | 1 |
| MATHEMATICS | | | 1 | | | 1 | | | 2 |
| MUSIC | | | 1 | 1 | | | | | 2 |
| PHYSICS | | | | | 2 | | | | 2 |
| PSYCHOLOGY | | | 2 | 1 | 2 | | | | 5 |
| SOCIOLOGY | | | 1 | | | | | | 1 |
| | 1 | 7 | 42 | 123 | 250 | 245 | 69 | 3 | 740 |

by their verbal and quantitative scores. The hierarchical clustering procedure alluded to above was also applied to this data.

5.2 Results

Inasmuch as the results of the clusterings using the two data sets were quite similar, we will illustrate our findings with the GRE Subject Test data. For this example, only information on the type of Subject Test taken (and not the score on the test) was employed. In hierarchical clustering, clusters, once formed, are never split but rather are combined with other clusters. In Figure 7, we display the last 13 steps of the algorithm, showing how the 14 groups formed by that stage are combined. For example, at step 13, chemistry is combined with the medical sciences group. This new cluster is then linked to a biological sciences group at step 8. The different disciplines contained within each of the 14 groups are listed in Table 5. The labels assigned to these groups are only meant to be suggestive, as there are a number of anomalies present. For example, mining appears in the paramedical group, well-separated from geology, while sociology is allied with French, Spanish and music.

It appears that these anomalies arise because of small sample sizes in some disciplines. On the whole, however, the clusters are quite sensible and provide some useful insight into the similarities of applicants to the various disciplines. Note that chemistry and physics are quite widely separated although they are both academic scientific disciplines while economics is somewhat removed from the quantitative sciences. It remains to be seen whether these clusters

Figure 7

Clusters of Graduate Fields. Numbers indicate number of steps before end of algorithm at which clusters are joined. Dashed lines denote five clusters employed in validity trials.

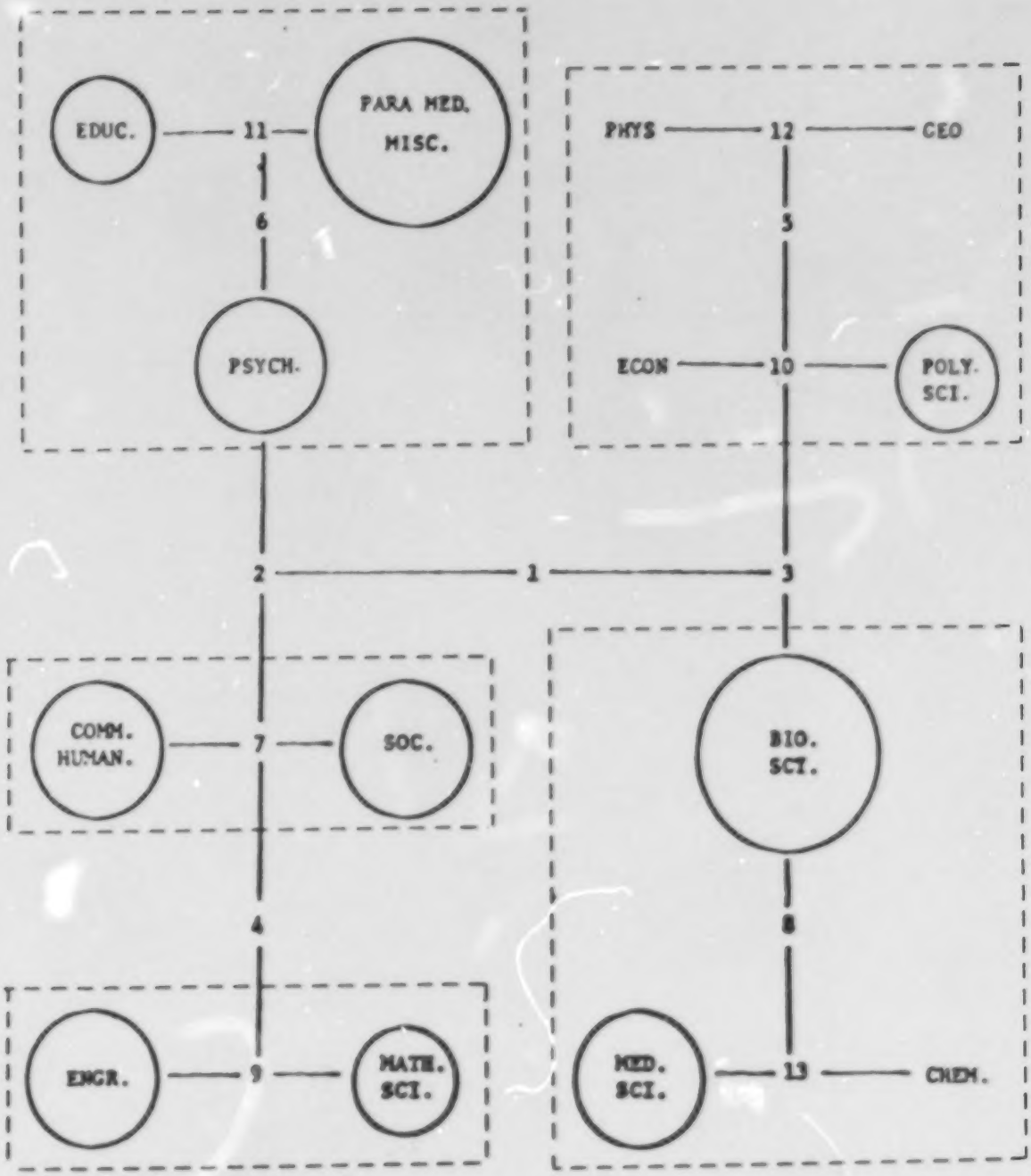


TABLE 5

Grouping of Disciplines for 14 Clusters in Figure 7

Education

Educational Administration
Education
Physical Education

Paramedical & Miscellaneous

Other
Speech
Hospital Administration
Nursing
Audiology
Occupational Therapy
Home Economics
Other
Architecture
Mining
Business and Commerce
Geography
Law
Urban Development

Psychology

Educational Psychology
Guidance and Counseling
Social Psychology
Anthropology
Other Social Sciences
Industrial Relations
Social Work
Psychology

Sociology

French
Spanish
Sociology
Music

Mathematics

Applied Mathematics
Statistics

Communications and Humanities

Linguistics
Other Foreign Languages
Russian
Fine Arts
Communications
Archeology
Classical Languages
Religion
Far Eastern Languages
Library Science
Journalism
Italian
Dramatic Arts
Other Humanities
Art History
American Studies
German
Comparative Literature
Philosophy
English

Chemistry

Chemistry

Biological Sciences

Genetics
Microbiology
Biology
Anatomy
Optometry
Bacteriology
Entomology
Veterinary Medicine
Botany
Zoology
Other Biological Sciences
Physiology
Dentistry
Pathology
Forestry
Parasitology
Physical Therapy
Public Health
Agriculture

TABLE 5 (contd.)

Mathematics (contd.)

Mathematics
Computer Science

Engineering

Aeronautical Engineering
Metallurgical Engineering
Electrical Engineering
Civil Engineering
Mechanical Engineering
Chemical Engineering
Industrial Engineering
Other Engineering

Medical Sciences

Pharmacology
Medicine
Nutrition
Biochemistry
Biophysics
Pharmacy
Oceanography
Other Physical Sciences

Physics

Astronomy
Physics

Geology

Geology

Political Science

Slavic Studies
International Relations
Public Administration
Political Science
Near Eastern Languages
History

Economics

Economics

will provide useful information on the flow of students into graduate school. Presumably, one would also have to study the structure of the common applicant pools.

5.3 Applications to Validity

The original purpose of this portion of the analysis was to produce new clusters to which we could apply models of the form EBCf. Therefore, we employed the final five clusters yielded by the algorithm and indicated on Figure 7. However, fitting EBCf to these five clusters did not produce departmental regression planes with characteristics superior to those of EBFf. We did not attempt to fit the model to the full 14 cluster partition displayed in Figure 7.

Our second approach to the problem was to construct a matrix of distances between disciplines based on the homogeneity of the distributions of counts in the Subject Test-taken/Test-score matrix associated with each discipline. The multidimensional scaling program, MDSCAL (Kruskal & Wish, 1978), was then applied to the distance matrix. Three-, four- and five-dimensional scalings were produced, but only the five-dimensional representation was judged minimally acceptable in capturing the salient features of the data.

The corresponding five coordinates for each discipline were then added to the departmental level covariates for all departments in the discipline. Thus, in the empirical Bayes framework we have associated with each department a number of department-specific covariates based on average test scores of the students and a number of discipline-specific covariates based on national applicant characteristics. An empirical Bayes model was then fitted to the data,

using no cluster information. Once again, the prediction planes generated for each department proved no better than those derived from EBFf, which employs no discipline-linked covariates. Thus, our conclusion is that while the applicant data and clustering methods we have considered may promise intriguing possibilities, more work is required before these possibilities are realized.

## 6. Conclusions

The statistical analyses we have carried out indicate that the values of the coefficients of a prediction equation for a department are strongly related to the typical test scores earned by students in that department. Moreover, this relationship appears to be a linear one and does not depend on the type of department. Empirical Bayes methods make use of this structure to obtain estimates of the coefficients that improve significantly upon those derived by least squares using individual department data.

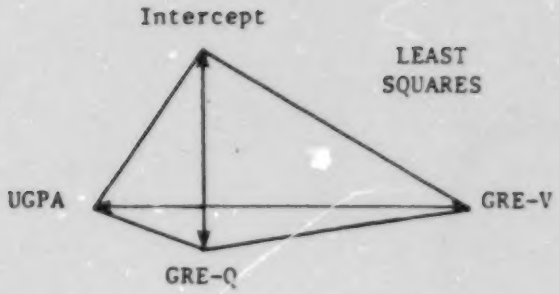
That different departments with similar students (in terms of predictor scores) have similar prediction equations is not a little surprising, given the variety of departments involved. What is more surprising, better departments appear generally to have stronger levels of association between criterion and predictors. Although we have pretty much ruled out differential restriction of range as a confounding factor, there are a number of plausible explanations. For example, grading standards may be more carefully observed, or academic ability more important for success, in departments with the more able students. It is an intriguing puzzle that demands the attention of specialists in higher education. While we cannot shed much light on this issue, we can make some comparisons between the results of the least squares and empirical Bayes methods.

Of necessity, the dispersion of the empirical Bayes coefficients across departments is much less than that for the least squares coefficients. The cross-validation analyses showed that the empirical Bayes

models produce better predictions for replicate data. Equally important, the empirical Bayes coefficients in the multiple regressions are essentially always positive and more stable across replicate data sets. To demonstrate the comparative stability of coefficients derived through empirical Bayes, Figure 8 was constructed. One department was selected at random from each of four major groups of disciplines: humanities, biological sciences, social sciences, and physical sciences. The empirical Bayes and least squares estimates based on the half-samples for each department were obtained, and the absolute value of the differences for each coefficient was computed. Thus, if the coefficients for one half-sample of a particular department were $(b_o, b_i, b_2, b_3)$ and for the other half-sample $(b_o', b_1', b_2', b_3')$, then the quantity computed was the vector $( b_o - b_o' , b_1 - b_1' , b_2 - b_2' , b_3 - b_3' )$.

This vector can be represented graphically in a number of ways. One such way is by means of an icon. Figure 8 illustrates the use of an asymmetric diamond icon. Essentially, each of the components is plotted as a vector starting at the origin and extending outward in one of the compass directions. The ends of these four vectors are connected to form a quadrilateral. The size and shape of the quadrilateral indicate both the absolute and relative size of the componentwise differences. As Figure 8 convincingly demonstrates, differences between the empirical Bayes estimates from one half-sample to another are much smaller than those for the least squares estimates: In each case, the empirical Bayes quadrilateral is entirely contained within the least squares quadrilateral.

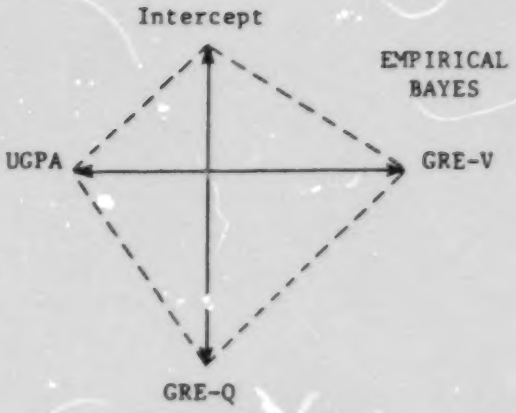
Figure 8




Figures 8a, b, c, and d:
(see following 2 pages)

Differences across half-samples in the estimates of the four regression coefficients in a prediction equation by two methods: Empirical Bayes and Least Squares.

Note different scales on abscissa and ordinate axes and across all four sample departments.



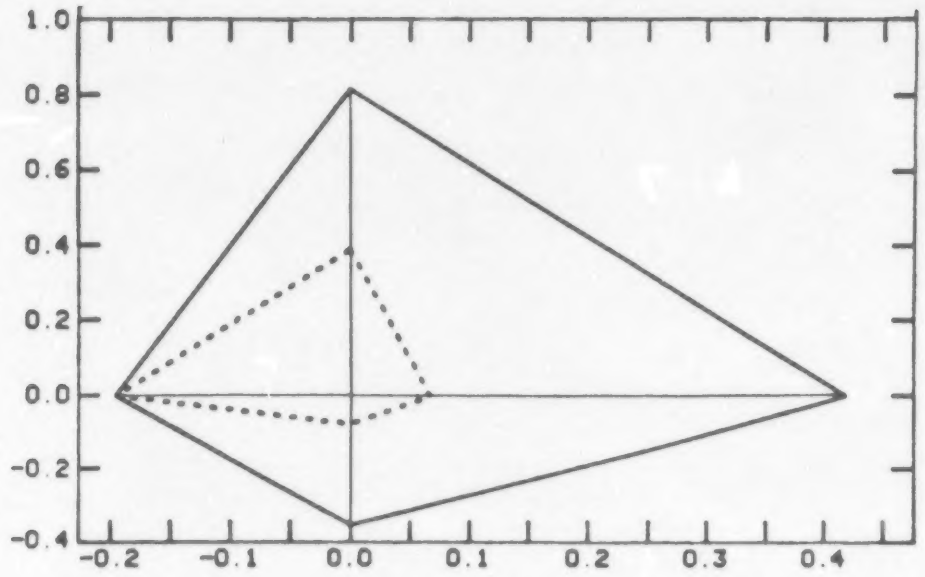



Figure 8a: Sample Physical Sciences Department

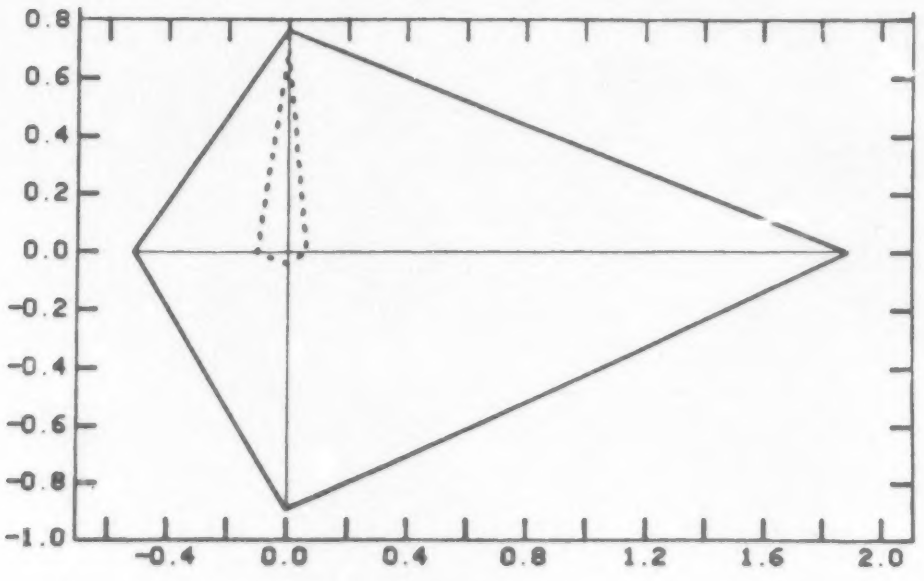



Figure 8b: Sample Biological Sciences Department

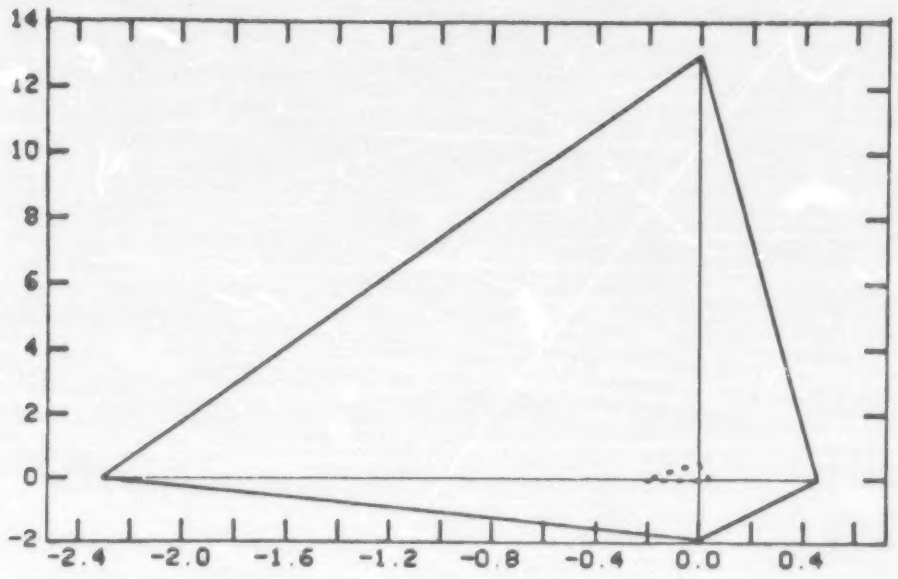



Figure 8c: Sample Humanities Department

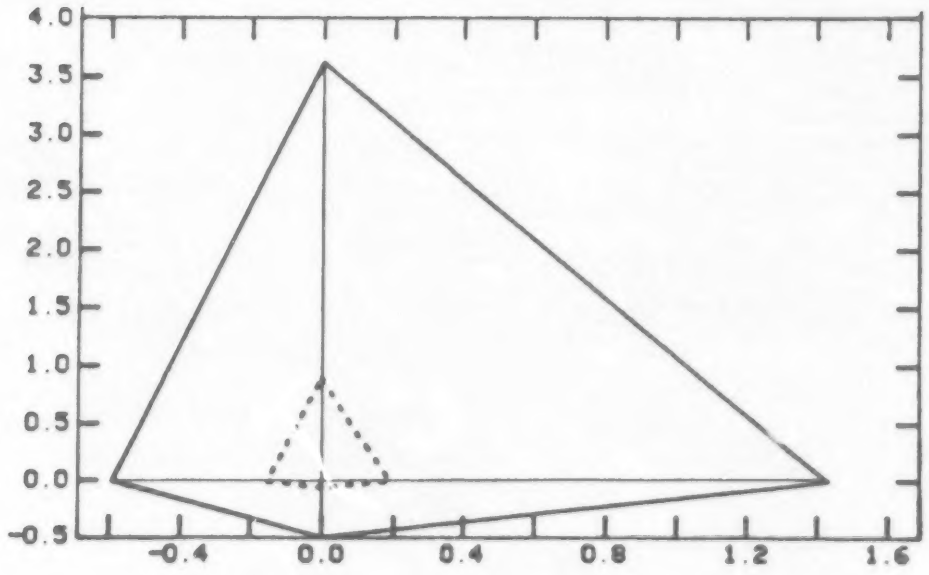



Figure 8d: Sample Social Sciences Department

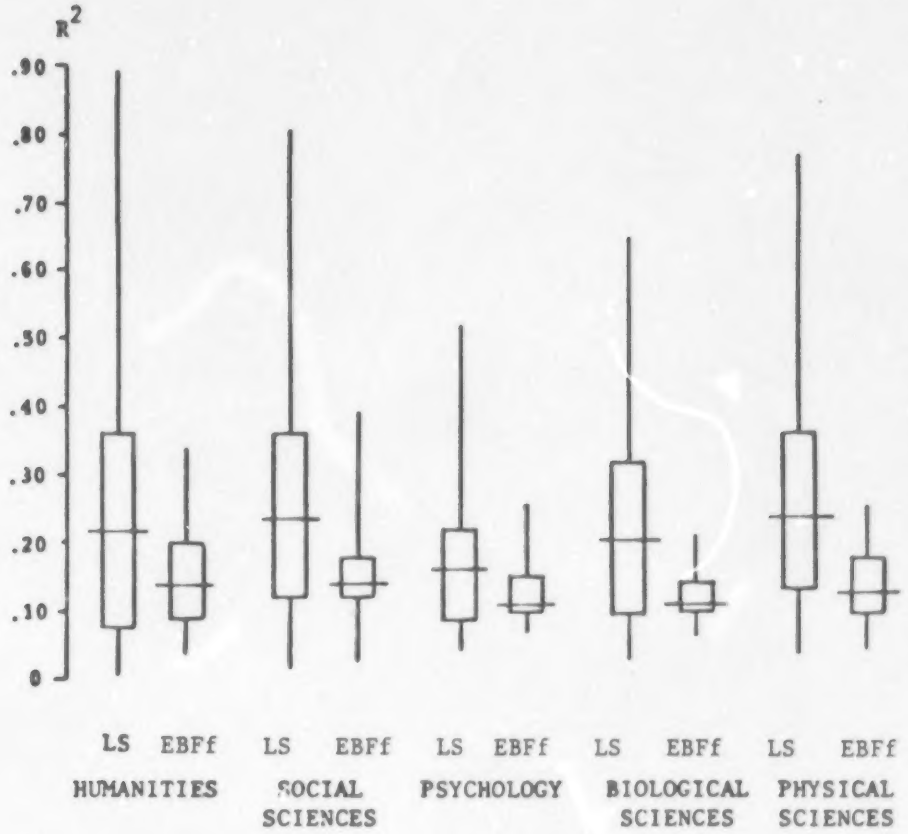
These properties are very important in the context of the admission process. First, admission committees are rightly disturbed by large year-to-year fluctuations in the prediction equations. Stability in the prediction equations is not only plausible but also facilitates the development of a consistent set of admission criteria. Second, the prediction equations are fitted to data derived from the most recent cohort to complete the first year of studies but are applied to prospective candidates. The predicted FYAs for these candidates may play a role in the admission decisions, and the extent of that role will depend on how confident the committee is that the predicted and actual FYAs will tend to be similar. One measure of the similarity is the $R^2$ coefficient, which, in this case, would be the proportion of variation in actual FYA explained by the FYAs generated by the fitted model.

The simplest approach is to take the $R^2$ of the least squares fit to the past cohort as an estimate of the $R^2$ that will be realized when the predictions generated by that same fit are compared to the actual FYAs of the new cohort. The cross-validations showed that such estimates tend to be biased and considerably overstate the $R^2$. On the other hand, the so-called predicted $R^2$ defined in Section 3, when applied to empirical Bayes estimates, provided reasonable and approximately unbiased estimates of the $R^2$ to be realized on new data with predictions generated from a regression derived from other data. Inasmuch as the predicted $R^2$ based on half-samples proved so useful, we have adapted it to the full samples.

Figure 9 compares the distribution of predicted $R^2$ for EBFf with the distribution of $R^2$ of the least squares fit for departments in the five clusters. Note that the $R^2$ distributions for empirical Bayes display considerably less dispersion. Our analysis strongly suggests that these $R^2$ values are more reliable in a prediction context than those of least squares. The differences by cluster in the distributions deserve further study. It should be emphasized that the cross-validation $R^2$ we discuss here is somewhat different in nature from the $R^2$ typically produced in validity studies. The latter is based on regressions fitted to the full data of the current year, while the former employs regression fitted to the previous year's data.

In more familiar terms, our results indicate that a typical department employing a prediction equation estimated by the methods introduced here should realize validities between 0.3 and 0.5 in predicting the grades of those candidates for admission who will eventually attend and persist. Because restriction of range corrections have not been implemented, the validities are rather lower than they would be in unselected samples. Recent work on the validity of the LSAT (Braun & Szatrowski, 1982) has demonstrated that prediction planes estimated through empirical Bayes, combined with test score information on applicants, can provide estimates of validity corrected for selection. These procedures are perfectly practicable in the context of the GRE Validity Study Service data, but have not been carried out here. The inclusion of scores on the Subject Tests

Figure 9: $R^2$ for Least Squares and EBFf Based on Full Samples



among the predictors would also enhance the validity of the prediction equation.

The employment of empirical Bayes methods also has facilitated study of the question of whether constructing separate prediction planes for particular groups of students leads to better predictions. We investigated two different ways of classifying students: One by age and sex and the other by race alone. In both cases, a single prediction plane for all students performed essentially as well as the set of separate prediction planes. Thus, within the constraints set by the quantity and quality of the data, there is no evidence of differential predictive validity for the common predictors of graduate school performance.

Finally, a methodological foray into cluster analysis has led to a new approach to the clustering of academic disciplines. Various attempts to capitalize on these clusters, in conjunction with empirical Bayes, to produce improved prediction schemes were unsuccessful. Nonetheless, we believe the methods we have developed should prove useful. In particular, these new clusters of disciplines may lead to some insight into shifts in the flow of students into various areas of graduate study.

What then are the implications of our work for the GRE Validity Study Service, and what directions should future research pursue? First, we believe that our finding of an apparently universal structure underlying the prediction planes for graduate departments is far-reaching. Not only is it an interesting result in a purely theoretical sense, demanding some explanation, but also it holds the

promise of providing the Validity Study Service with an easily implementable prediction system that can significantly widen the applicability and reliability of the annual validity studies. Almost all graduate departments can be provided with useful and replicable results concerning the roles that various predictors, singly or jointly, can play in the prediction of the future performance of prospective students.

In this report, we have stressed methodological considerations and general conclusions over results for specific departments. This follows from the fact that we have only begun to explore the rich family of models within the empirical Bayes framework. Further work will undoubtedly uncover other models that share the conceptual basis of those we have described here, but lead to different sets of departmental coefficients that perform better in practice. The work we have carried out can lead directly to a significant increase in the number of departments participating in the Validity Study Service as well as enhance the quality and breadth of information the validity reports can provide the individual department.

A first priority of future research should be to continue the development of practical empirical Bayes models and to test them in the crucible of cross-validation against more classic systems involving various levels of data pooling. Once an acceptable system has been developed, a comprehensive examination of the resulting prediction equations should be undertaken. This should include a study of the incremental contributions to validity of each predictor, once the other predictors are accounted for. At that point, a qualitative

analysis of the contributions of the various predictors by discipline should be carried out so that departments that do not employ formal prediction systems could still benefit from the results.

In addition, two other problems must be addressed. One concerns the use of unique predictors by individual departments. The empirical Bayes framework must be suitably expanded to include such predictors, while maintaining the desired stability in the final prediction equations. Second, the question of differential validity for various subgroups of candidates must be attacked in a somewhat different fashion. It is possible that a hybrid of various models can be formulated to overcome the paucity of data on minorities.

The empirical Bayes framework is undoubtedly rich enough to accommodate even this difficulty. Our work suggests that the GRE Validity Study Service will benefit from incorporating empirical Bayes ideas. Although the models we have experimented with have quite reasonable properties, we are convinced that further research will uncover still more powerful ways of looking at validity data.

References

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics. New York: John Wiley.

Boldt, R. R. (1975). Comparison of a Bayesian and a least squares method of educational prediction. GRE Board Professional Report No. 70-3P. Princeton, N.J.: Educational Testing Service.

BMD-P-77. (1977). Biomedical computer programs P-series. Edited by W. J. Dixon. Los Angeles, Calif.: University of California Press.

Braun, H. I., & Jones, D. H. (1981). The Graduate Management Admission Test: Prediction bias study. GMAC Research Report 81-4, RR-81-25. Princeton, N.J.: Educational Testing Service.

Braun, H. I., & Jones, D. H. (1982). Hierarchical clustering of nominal data. Proceedings of the American Statistical Association, Section on Social Statistics.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model, from data of deficient rank. Psychometrika, 48, No. 2, 171-181.

Braun, H. I., & Szatrowski, T. H. (1982). Development of a universal grade scale for American law schools and the reconstruction of ideal validity experiments. Draft Final Report. Educational Testing Service: Princeton, N.J.

Burton, N. (1982). Private communication.

Dawes, R. M. (1975). Graduate admission variables and future success. Science, 187, 721-3.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (B) 39, 1-38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. Journal of the American Statistical Association, 76, 341-353.

Hartigan, J. A. (1975). Clustering algorithms. New York: John Wiley and Sons.

Kruskal, J. B., & Wish, M. (1978). Multidimensional scaling. Beverly Hills, Calif.: Sage Publications.

Lindley, D. V. (1969). A Bayesian solution for some educational prediction problems, II. Research Bulletin 69-71. Princeton, N.J.: Educational Testing Service.

Lindley, D. V. (1970). A Bayesian solution for some educational prediction problems, III. Research Bulletin 70-33. Princeton, N.J.: Educational Testing Service.

Livingston, S. A., & Turner, N. J. (1982). Effectiveness of the Graduate Record Examinations for predicting first-year grades: 1980-81 summary report of the Graduate Record Examinations Validity Study Service. Princeton, N.J.: Educational Testing Service.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. (1972). Estimating multiple regressions in m-groups: A cross-validation study. British Journal of Mathematical and Statistical Psychology, 25, 33-50.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association, 75, 801-816.

Stone, M. (1978). Cross-validation: A review. Mathematische Operationsforschung und Statistik, 9, 127-140.

Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Addison-Wesley.

Wilson, K. M. (1982). A study of the validity of the restructured GRE Aptitude Test for predicting first-year performance in graduate study. GRE Board Research Report No. 78-6R. Princeton, N.J.: Educational Testing Service.

Wilson, K. M. (1982). Private communication.

Wilson, K. M. (1979). The validation of GRE scores as predictors of first-year performance in graduate study: Report of the GRE Cooperative Validity Study Project. GRE Board Research Report No. 75-8R. Princeton, N.J.: Educational Testing Service.

Technical Appendix

The estimation problem following from the empirical Bayes model may be described as follows. Suppose there are m departments with $n_i$ students in department i. We assume that a linear model of the following form holds:

(1) $Y_i = X_i B_i + e_i \qquad (i = 1, 2, \ldots, m)$

where

$Y_i$ is an $n_i$ x 1 vector of first-year averages,

$X_i$ is an $n_i$ x P design matrix containing information on P test scores of the students,

$B_i$ is a P x 1 vector of regression coefficients, and

$e_i$ is an $n_i$ x 1 vector of random errors.

$Y_i$ and $X_i$ are observed, while $B_i$ and $e_i$ are not. Interest centers on obtaining estimates of $\{B_i\}$. Ordinarily, one assumes that the components of $e_i$ are a ramdom sample from a normal distribution, $N(0, \sigma_i^2)$. The key assumption in our formulation of an emprirical Bayes model is that

(2) $B = Z G + D$

where B is an m x P matrix of regression coefficients where the ith row of B is denoted $B'_i$,

Z is an m x k matrix of departmental covariates where the ith row of Z is denoted $Z'_i$ and contains information on k variates,

G is a k x P matrix of regression coefficients,

D is an m x P matrix of random errors where the ith row of D is denoted $D'_i$.

It is assumed that the $D'_i$ values are distributed according to a multivariate normal distribution, $N(0, \Sigma^*)$.

Under the models (1) and (2), together with the accompanying normality assumptions, the $\{B_i\}$ estimates have a joint multivariate normal distribution given the observed data and the parameters G and $\Sigma^*$. The empirical Bayes estimate of the $\{B_i\}$ is the vector of means of this posterior distribution with G and $\Sigma^*$ replaced by their maximum likelihood estimates. These estimates are obtained through application of the EM algorithm (Dempster, Laird, & Rubin, 1977). The algorithm consists of a succession of two-step cycles: an E-step and an M-step.

Beginning with initial estimates of $\sigma^2_i$, G, and $\Sigma^*$, we obtain for the E-step:

$$r_i = E(B_i | Y_i, X_i, \hat{\sigma}^2_i, \hat{G}, \Sigma^*)$$

$$= (P^* + \hat{P}_i)^{-1} \quad P^*\hat{G}' Z_i + \hat{P}_i\hat{B}_i$$

and

$$S_i = E(B_iB_i' | Y_i, X_i, \hat{\sigma}_i^2, \hat{G}, \hat{\Sigma}^*)$$

$$= r_ir_i' + (P^* + \hat{P}_i)^{-1}$$

where $P^* = (\Sigma^*)^{-1}$, $\hat{P}_i = \hat{\sigma}_i^{-2} X_i'X_i$, and $\hat{B}_i$ is the least squares estimate of $B_i$, based on data from department i only. Thus, the current estimate of the posterior mean of $B_i$ is a precision-weighted combination of the least squares estimate and the appropriate point on the surface defined by equation (2). The quantity $S_i$ is required for the M-step.

In the M-step, we obtain updated estimates of $\sigma^2_i$, G and $\Sigma^*$ based on $\{r_i\}$ and $\{S_i\}$. In effect, we regress the $r_i$ (in place of the

unobservable $B_i$) on $Z_i$. Thus

$$\hat{G} = (Z'Z)^{-1} Z'R$$

$$\hat{\Sigma} = m^{-1} \; B'B - \hat{G}'(Z'Z)\hat{G}$$

and

$$\hat{\sigma}^2_i = n_i^{-1} \; Y_i'Y_i - 2r'_i X_i'Y_i + \sum_{j,k} s_{jk}^{(i)} \, w_{jk}^{(i)} \; ,$$

where $S_i = (s_{jk}^{(i)})$, $X_i'X_i = (w_{jk}^{(i)})$, and R is a matrix whose ith row is $r_i'$.

The E-step is then reentered with the updated estimates, and new conditional expectations are calculated. The process is continued until the estimates of G, $\Sigma^*$, and $\{\sigma_i^2\}$ converge to the maximum likelihood estimates. The corresponding $\{r_i\}$ values are the desired empirical Bayes estimates of $\{B_i\}$.

It should be noted that we do not attempt to obtain simultaneously empirical Bayes estimates of $\{B_i\}$ and $\{\sigma_i^2\}$. This more complicated problem cannot be solved directly, and extensive numerical calculations are required to obtain even posterior modes as estimates of the parameters. The quality of such estimates is not clear. We prefer, therefore, to uncouple the estimation problems for the residual variances and the regression coefficients.